

**„Gefühl ist alles; Name ist Schall und Rauch.“ – Der Einsatz von Sentiment Analysis in der quantitativen Dramenanalyse**

Masterarbeit im Fach Medieninformatik am

Institut für Information und Medien, Sprache und Kultur (I:IMSK)

Vorgelegt von: Thomas Schmidt

Adresse: Thannsteinweg 10, 93049 Regensburg

Matrikelnummer: 1441397

Erstgutachter: Prof. Dr. Christian Wolff

Zweitgutachter: Dr. Manuel Burghardt

Laufendes Semester: SoSe 2017

Abgegeben am: 09.10.2017

Inhalt

[1 Einleitung 7](#_Toc494151864)

[2 Related Work 7](#_Toc494151865)

[2.1 Sentiment Analysis – Grundlagen 7](#_Toc494151866)

[2.2 Lexikon-basierte Ansätze 15](#_Toc494151867)

[2.3 Deutschsprachige Sentiment-Lexika 22](#_Toc494151868)

[2.3.1 SentimentWortschatz – SentiWS 22](#_Toc494151869)

[2.3.2 Berlin Affective Word List – Reloaded (BAWL-R) 24](#_Toc494151870)

[2.3.3 NRC Word-Emotion Association Lexicon (NRC) 26](#_Toc494151871)

[2.3.4 Clematide-Dictionary (CD) 28](#_Toc494151872)

[2.3.5 German Polarity Clues (GPC) 29](#_Toc494151873)

[2.3.6 Sonstige SA-Lexika 31](#_Toc494151874)

[2.4 Sentiment Analysis in der Literaturwissenschaft 34](#_Toc494151875)

[2.5 Quantitative Dramenanalyse 45](#_Toc494151876)

[3 Forschungsfragen und -Agenda 46](#_Toc494151877)

[4 Dramen-Korpus 46](#_Toc494151878)

[5 Back End: Sentiment Analysis 46](#_Toc494151879)

[5.1 Konzeption 46](#_Toc494151880)

[5.2 Vorverarbeitung der Dramen 46](#_Toc494151881)

[5.2.1 Sprachverarbeitung – Lemmatisierung 46](#_Toc494151882)

[5.2.1.1 Idee 46](#_Toc494151883)

[5.2.1.2 Entwicklung 46](#_Toc494151884)

[5.3 Lexika-Verarbeitung 46](#_Toc494151885)

[5.3.1 Grundsätzliche Verarbeitung 46](#_Toc494151886)

[5.3.2 Lexika-Kombination 46](#_Toc494151887)

[5.3.3 DTA – Erweiterung 46](#_Toc494151888)

[5.3.3.1 Idee 46](#_Toc494151889)

[5.3.3.2 Entwicklung 46](#_Toc494151890)

[5.3.4 Sprachverarbeitung - Lemmatisierung 46](#_Toc494151891)

[5.3.4.1 Idee 46](#_Toc494151892)

[5.3.4.2 Entwicklung 46](#_Toc494151893)

[5.3.5 Stoppwortlisten 46](#_Toc494151894)

[5.3.5.1 Idee 46](#_Toc494151895)

[5.3.5.2 Entwicklung 46](#_Toc494151896)

[5.4 Vorverarbeitung – Überblick 46](#_Toc494151897)

[5.5 Sentiment Analysis 46](#_Toc494151898)

[5.5.1 Konzeption 47](#_Toc494151899)

[5.5.2 Entwicklung 47](#_Toc494151900)

[5.5.3 SA-Metriken 47](#_Toc494151901)

[6 Lexika-Analyse 48](#_Toc494151902)

[6.1 Idee 48](#_Toc494151903)

[6.2 Entwicklung 48](#_Toc494151904)

[6.3 Ergebnisse 48](#_Toc494151905)

[6.4 Diskussion 48](#_Toc494151906)

[7 Vokabular-basierte Evaluation 48](#_Toc494151907)

[7.1 Idee 48](#_Toc494151908)

[7.2 Entwicklung 48](#_Toc494151909)

[7.3 Ergebnisse 48](#_Toc494151910)

[7.4 Diskussion 48](#_Toc494151911)

[8 Test-Korpus-Erstellung 48](#_Toc494151912)

[9 Test-Korpus-Evaluation 48](#_Toc494151913)

[9.1 Idee 48](#_Toc494151914)

[9.2 Entwicklung 48](#_Toc494151915)

[9.3 Ergebnisse 48](#_Toc494151916)

[9.4 Diskussion 48](#_Toc494151917)

[10 Front-End – Visualisierung 48](#_Toc494151918)

[10.1 Konzeption 48](#_Toc494151919)

[10.2 Entwicklung 48](#_Toc494151920)

[10.3 Funktionalität 48](#_Toc494151921)

[11 Fallbeispiele 49](#_Toc494151922)

[12 Diskussion 49](#_Toc494151923)

[Anhang A – B5T-Fragebogen 56](#_Toc494151924)

Abbildungen

**Es konnten keine Einträge für ein Abbildungsverzeichnis gefunden werden.**

Tabellen

Zusammenfassung

Abstract

# Einleitung

# Related Work

Im Folgenden Abschnitt wird die für das Projekt relevante Literatur beschrieben. Zunächst wird der Begriff der Sentiment Analysis definiert und grundsätzliche Konzepte erläutert. Dabei wird ein Schwerpunkt auf in der vorliegenden Arbeit verwendete Lexikon-basierte Verfahren gelegt und zentrale deutsche Sentiment-Lexika beschrieben. Abschließend werden analoge Projekte illustriert, die ebenfalls versuchen SA-Methoden und Literaturwissenschaft zu verknüpfen.

## Sentiment Analysis – Grundlagen

Sentiment Analysis (SA), häufig auch Opinion Mining oder Sentiment Detection genannt, wird von Liu (2016, S. 1) in seinem Standardwerk der Sentiment Analysis als Studienfeld definiert, dass sich mit der Analyse der Meinungen, Gefühle, Bewertungen, Einstellungen und Emotionen von Personen gegenüber Entitäten und dem Ausdruck der genannten Attribute in geschriebenem Text befasst. Das Ziel von Sentiment Analysis ist es mit Hilfe von computergestützten Methoden Gefühle und Meinungen aus natürlichsprachlichen Text zu extrahieren und Aussagen darüber zu treffen.

Die Motivation für dieses Forschungsfeld resultiert aus dem Umstand, dass Gefühle, Meinungen und Emotionen essentiell für fast alle menschlichen Aktivitäten sind (Liu, 2016, S. 3). Gefühle und Meinungen sind beispielsweise bedeutend für Regierungsorganisationen, um die öffentliche Meinung bezüglich ihrer Politik einzuschätzen (Liu, 2016, S. 4-7). Unternehmen möchten Kenntnisse über die Einschätzung zu den eigenen Produkten und Dienstleistungen erlangen (Liu, 2016. S. 4). Der individuelle Konsument nutzt nachweislich die Meinungen von Autoriäten und anderen Konsumenten um Kauf- oder Politik-Entscheidungen zu treffen und analysiert beispielsweise dazu Produkt-Reviews oder Mitteilungen in sozialen Medien (comScore, 2007; Horrigan, 2008; Rainie & Horrigan, 2007; Pang & Lee, 2008).

Liu (2006, S. 16-46) erstellt ein konzeptionelles SA-Framework basierend auf der Definition und Relation von verschiedenen theoretischen Konzepten. Zentrale Bestandteile einer Meinung (opinion) sind dabei der Meinungshalter, das Sentiment-Ziel und die Meinung bzw. das Gefühl des Meinungshalters gegenüber dem Sentiment-Ziel. Das Sentiment, definiert als Gefühl, Einstellung oder Emotion, die mit einer Meinung zusammenhängt, kann dabei unterschiedliche Ausprägungen annehmen. Bei Liu (2016) als auch meist in der Literatur kann das Sentiment positiv, negativ oder neutral sein, basierend auf der Evaluation des Meinungshalters gegenüber dem Ziel. Eine derartige Einteilung wird auch Polarität, semantische Orientierung oder Valenz genannt. Die einzelnen Klassen können auch verschiedene Intensitäten annehmen, also mehr oder weniger negativ oder positiv sein. Aufgabe der SA ist die konkrete computergestützte Feststellung des Sentiments. Tsytsarau und Palpanas (2012) greifen die Konzepte auf und definieren als zentrale Bestandteile der SA das Dokument das untersucht wird, das Thema über das eine Meinung oder ein Gefühl ausgedrück wird und wiederum die Form und Ausprägung dieser Meinung oder des Gefühls. Sie weisen auf die Bedeutungsdifferenzen von Meinungen, Gefühlen und Emotionen hin und konstatieren beispielsweise, dass Gefühle und Emotionen oft kein direktes Sentiment-Ziel besitzen. In ihrer Literaturanalyse stellen sie ferner fest, dass die große Mehrzahl der SA-Forschung Sentiment als binäres Konzept mit den Klassen positiv und negativ auffasst. Andere diskrete Erweiterungen, z.B. mit der Klasse neutral treten deutlich seltener auf. Kontinuierliche Ausprägungen, z.B. die Angabe von Polaritätsintensitäten über metrische Zahlenwerte sind ebenfalls im Vergleich eher selten. Die Verwendung und Prädiktion komplexer emotionaler Kategorien, beispielsweise eine Einteilung von Emotionen z.B. nach Wut, Ekel, Furcht, Freude, Traurigkeit, Überraschung wie Ergebnisse aus der Psychologie nahelegen (Ekmann et al., 1987), findet kaum Anwendung in der SA-Forschung (Tsytsarau und Palpanas, 2012). Das Polaritäts-Paradigma, also die Einteilung in positiv und negativ ist vorherrschend.

Sentiment Analysis ist auch eines der aktivsten Forschungsfelder in den Computerwissenschaften seit 2000 (Vinodhini & Chandrasekaran, 2012). Mittlerweile wurden über 7000 Artikel zu dem Themenfeld verfasst (Feldman, 2013). Die aktive Forschung hängt eng mit dem Aufkommen und der Popularität des World Wide Webs, sozialer Medien und des partizipativer Web-Plattformen zusammen, da somit große und für die SA geeignete Text-Ressourcen leicht verfügbar wurden (Feldman, 2013; Vinodhini & Chandrasekaran, 2012; Ravi, 2015; Liu, 2016, S. 3). Zentrale Plattformen, die für die Erhebung von Datenquellen genutzt werden, sind vor allem Review-Seiten wie Amazon (Bhatt et al., 2015; Fang & Zhan, 2015), IMDB (Film-Reviews; Mudinas et al., 2012) oder Rotten Tomatoes (Pang & Lee, 2005) aber auch Blogs (Godbole, Srinivasaiah & Skiena; 2007; Melville, Gryc & Lawrence, 2009). Eine weitere häufig genutzte Datenquelle sind Microblogging- und Social-Media-Plattformen, vor allem Twitter (Pak & Paroubek, 2010; Kouloumpis, Wilson & Moore, 2011).

Der Einsatz von Sentiment Analysis wird dabei auf zahlreichen unterschiedlichen Anwendungsfeldern untersucht. Zentrale Gebiete sind beispielsweise die Analyse von Film-Reviews und die damit zusammenhängende Erfolgsvorhersage (Mishne & Glance, 2006; Liu et al., 2007; Singh et al., 2009) sowie die Analyse von Produkten und Verkäufern auf Basis von Produkt-Reviews (Cui, Mittal & Datar, 2006; McGlohan, Glance & Reiter, 2010). Ein weiteres wichtiges Anwendungsfeld ist der Bereich der Politik. O‘ Connor et al. (2010) können beispielsweise einen Zusammenhang zwischen politischem Erfolg und der Sentiment-Bewertung auf Twitter nachweisen. Weitere SA-Forschung auf politischer Ebene findet man wieder über Tweets bei Tumasjan et al. (2010), mit News-Artikeln bei Khoo et al. (2012) oder auch bezogen auf die letzten Präsidentschaftswahlen der USA wieder über Tweet-Analyse bei Chin, Zappone und Zhao (2016). Ein weiterer Forschungsschwerpunkt der Sentiment Analysis ist die Analyse von Artikeln auf News- oder Zeitungs-Webseiten (Balahur et al., 2009; Balahur et al., 2013) oder von News-Feeds (Wanner et al., 2009) um Sentiment-Aussagen über erwähnte Entitäten oder Themen der Artikel zu machen. Zueletzt sei noch die Analyse von Börsen- und Marktnachrichten gennant um mit Hilfe der SA Prädiktionen über Erfolgsentwicklungen zu machen (Das & Chen, 2007; Zhang, Fuehrers & Gloor, 2011, Bollen, Mao & Zeng, 2011).

Innerhalb der Forschung und der praktischen Umsetzung von SA gibt es verschieden Ansätze, die verfolgt werden und unterschiedliche Ebenen auf denen die SA ausgeführt werden kann. Liu (2016) unterscheidet zwischen einer Dokument-, Satz- und Aspekt-Ebene. Auf Dokument-Ebene wird die Gesamt-Meinung bzw. das Gesamt-Sentiment eines ganzen Dokuments klassifiziert, also beispielsweise eine einzelne Produkt-Review, und ob diese eine insgesamt positive oder negative Einschätzung bezüglich dem beschriebenen Produkt ausdrückt. Das gleiche wird mit einzelnen Sätzen auf der Satz-Ebene ausgeführt. Auf der Aspekt-Ebene wird nun versucht, das explizite Ziel einer Meinung oder eines Sentiments in einem Text zu ermitteln. Neben der Ebenen-Taxonomie von Liu (2016) findet man insbesondere im Forschungsfeld von Sentiment-Lexika (siehe Kapitel XX) noch die Wortebene als unterste Form der Sentiment-Bestimmung (Esuli & Sebastiani, 2006; Missen, Boughanem & Cabanac, 2013). Hier wird über verschiedene Methoden beispielsweise versucht die Polarität eines einzelnen Wortes zu klassifizieren. Es finden sich noch weitere Betrachtungsebenen in der Literatur, die manchmal eng mit den genannten Konzepten zusammenhängen, als weitere Beispiele seien die Phrasen-Ebene (Tan et al., 2012) oder auch die Vergleichs-Ebene (Feldman, 2013) genannt. Bei letztgenannter Ebene steht der Sentiment-basierte Vergleich von mehreren Entitäten im Vordergrund, welche beispielsweise in einem Text präferiert werden (Feldman, 2013). Tsai et al. (2013), Mudinas, Zhang & Levene (2012) oder auch Cambria et al. (2013) befassen sich mit einer „Konzept-Ebene“.

Bezüglich des methodischen Vorgehens gibt es verschiedene Klassifikkationsansätze in unterschiedlichen Meta-Studien. Konsistent findet man jedoch stets eine Unterteilung in folgende zwei Gruppen: Verfahren des maschinellen Lernens und Lexikon-basierte Methoden (Vinodhini & Chandrasekaran, 2012; Medhat, Hassan & Korashy, 2013; Collomb et al., 2014; Tsytsarau & Palpanas, 2012; Kaur & Gupta, 2013; Ravi, 2015; D‘ Andrea et al., 2015).

Die Gruppe des Maschinellen Lernens fasst alle Ansätze zusammen bei denen mit Hilfe von Lernalgorithmen versucht wird anhand eines mit Sentiment-Informationen ausgezeichneten Datenkorpus einen Klassifikationsalgorithmus zu trainieren (Collomb et al., 2014). Das grundsätzliche Vorgehen wird beispielhaft anhand Pang, Lee und Vaithyanathan (2002) geschildert: Zunächst wird ein Korpus von Film-Reviews aus der Plattform IMDB bezogen. Die einzelnen Reviews enthalten dabei entweder eine numerische Bewertung oder eine klassische Bewertung nach einem 1-5 – Sterne-System, so dass sie einfach in positive oder negative Bewertungen unterteilt werden können. Auf diese Weise kann ein annotierter Korpus von ca. 2000 Bewertungen erstellt werden. Dieser Korpus wird als Trainingskorpus für drei Standard-Trainings-Algorithmen verwendet. Es wird ein standardisiertes bag-of-features Framework verwendet, mit der Häufigkeite einzelner Wörtern und n-Grammen als Merkmalsvektoren eines Dokuments. Es werden die drei Standard-Algorithmen Naive Bayes, Maximum-Entropie-Methodik und Support Vector Machiness auf diese Weise getestet. Den jeweiligen mit dem Trainingskorpus trainierten Algorithmen können dann neue Film-Reviews übergeben werden, die dann je nach Auftreten von Wörtern und N-Grammen die eher in positiven oder negativen Reviews vorkommen bezüglich der Polarität als positiv oder negativ klassifiziert werden. Über eine Evaluation auf einem Ausschnitt des beschriebenen Korpus können auf diese Weise Klassifikationsgenauigkeiten von ca. 80% festgestellt werden. Ravi (2015) kann durch Analyse der Literatur feststellen, dass Maschinelles Lernen die am häufigsten eingesetzte Methodik in der Forschung zur Sentiment Analysis ist. Die beliebtesten Algorithmen sind dabei Support Vector Machines und Naive Bayes, andere Verfahren sind bislang eher selten. Es ist vor allem naheliegend derartige Methoden zu verwenden wenn man Zugriff auf eine angemessen große mit Sentiment-Informationen annotierte Datenquelle hat. Dies kann explizit, wie im oben beschriebenen Fall über positive oder negative Ratings vorliegen oder implizit, zum Beispiel indem der emotionale Ausdruck von Emoticons (Read, 2005; Chin et al., 2016) oder die Bedeutung von Hashtags auf Twitter (Davidov, Tsur & Rapopport, 2010a; Davidov, Tsur & Rapopport, 2010b) als Klasse eines Dokuments genutzt werden. Vorteil dieser Methode ist die Möglichkeit der Erzeugen von Vorhersagemodellen für sehr spezifische Anwendungsgebiete und die Anpassung an den speziellen Wortschatz eines Untersuchungsgegenstands. Nachteil ist einerseits die schlechte Übertragung von trainierten Modellen auf andere Anwendungsgebiete und die Notwendigkeit eines großen korrekt annotierten Trainingskorpus, dessen Akquirierung oft sehr kostenintensiv ist oder für manche Gebiete gar nicht möglich (D‘ Andrea et al., 2015).

Bei Lexikon-basierten Verfahren wird auf sogenannte Sentiment-Lexika zurückgegriffen. Dabei handelt es sich um Wort-Listen, die für jedes Wort einen Sentiment-Score basierend auf der semantischen Orientierung des Wortes angeben, also z.B. ob das Wort positiv, negativ oder neutral ist. Durch simple Kalkulation wird auf Basis des Vorkommens dieser Wörter in einem Dokument oder Satz, die Sentiment-Ausrichtung des Dokuments oder Satzes bestimmt (Collomb et al., 2014; D‘ Andrea et al., 2015; Liu, 2016; S. 10-11). In der vorliegenden Arbeit werden Lexikon-basierte Verfahren als SA-Lösung aufgegriffen. Aus diesem Grund werden die Erstellung von Lexika, der Einsatz dieser, Vor- und Nachteile sowie die wichtigsten SA-Lexika in den nachfolgenden Kapiteln im Detail besprochen.

Neben Methoden maschinellen Lernens und Lexikon-basierten Verfahren existieren noch weitere Ansätze. Collomb et al. (2014) benennen noch als dritten Ansatz Regel-basierte Ansätze. Regel-basierte Ansätze werden jedoch meist als Erweiterung von Lexikon-basierten Verfahren (Ding, Liu & Yu, 2008; Asghar et al., 2017) oder in hybriden Modellen (Choi & Cardie, 2008; Balage Fhilo & Pardo, 2013) genutzt. In Kombination mit den Sentiment-Informationen eines Sentiments-Lexikons auf Worte-Ebene werden beispielsweise Regeln bezüglich Negationen, „Booster Words“ (auch Intensifier genannt), Idiomen oder Emoticons aufgestellt, um die Klassifikation zu verbessern. Ashgar et al. (2017) nutzen eine Liste von Modifier-Wörtern, also Wörtern die das Sentiment eines nachfolgenden Wortes verstärken oder verringern um den einzelnen Sentiment-Score zu präzisieren. Ähnlich gehen sie mit Negationen um, indem das Sentiment umgedreht wird. Eine anderes Beispiel für den Einsatz von Regeln, ist die Zuweisung als positiv wenn ein Tweet überwiegend positive Emoticons enthält (Balage Fhilo & Pardo, 2013). Ashgar et al. (2017) können mit Hilfe von Regeln eine Verbesserung bezüglich der Klassifikationsleistung im Vergleich zur herkömmlichen Nutzung eines Sentiment-Lexikons feststellen.

D‘ Andrea et al. (2015), Ravi (2015) und Medhat et al. (2014) identifizieren noch die Gruppe hybrider Lösungen. Darunter werden meist alle Ansätze verstanden, die verschiedene der obigen Methoden kombinieren, um Sentiment Analysis durchzuführen, so zum Beispiel die Kombination von maschinellen Lernen und der Verwendung von SA-Lexika. Ein Ansatz ist dabei beispielsweise ein nicht annotiertes Dokument-Korpus mittels den Ergebnissen eines Lexikon-basierten Sentiment-Berechnung zu klassifizieren und das Korpus mit den erhaltenen Klassifikationen als Trainingsset für einen Lernalgorithmus zu verwenden (Sommar & Wielondek, 2015; Lalji & Deshmuk, 2016). Über eine ähnliche Hybrid-Methodik erzeugen Pak und Paroubek (2010) Pseude-Dokumente aus SA-Lexika um einen Lernalgorithmus zu trainieren. Ferner sei noch ein Ansatz von Balage Fhilo und Pardo (2013) als exemplarisches Beispiel skizziert. Sie stellen eine schrittweise Methoden-Pipeline zur Bestimmung der Polarität auf. Im ersten Schritt werden Emoticon-basierte Regeln genutzt, im zweiten Schritt die Kalkulation mittels einem SA-Lexikon und im dritten Schritt eine mit annotierten Beispielen trainierte SVM. Wird für die Ergebnisse einer Ebene ein gewisser Bestimmungsschwellenwert nicht erreicht, wird das Dokument auf die nächste Ebene übertragen und so weiter bis das Sentiment in einer Ebene gemäß Schwellenwert eindeutig bestimmt wird. Appel et al. (2016) nutzen semantische Regeln, Fuzzylogik, Sentiment-Lexika und unüberwachtes Lernen für einen fortgeschrittenen und komplexen Hybrid-Ansatz. Weitere hybride Ansätze werden bei Ravi (2015), Collomb et al. (2014) und Thakkar und Patel (2015) zusammengefasst und erläutert.

Vergleicht man die Nutzung aller beschriebenen Methodengruppen stellen sowohl Tsytsarau und Palpanas (2012) als auch Ravi (2015) eine Dominanz von Sentiment Analysis basierend auf maschinellem Lernen fest. Dies liegt vor allem daran, dass die sich die Mehrzahl der Forschung auf Produkt- oder Film-Reviews fokussiert und es für diese Bereiche sehr einfach ist annotierte Trainingskorpora zu akquirieren, um themen-spezifisch optimierte Algorithmen zu trainieren. Medhat et al. (2014) stellen jedoch einen vermehrten Einsatz von SA-Lexika seit 2010 fest. Grund hierfür ist die einfache und generalisierte Nutzung und die Neuorientierung auf Themengebiete für die keine Trainingskorpora zur Verfügung stehen. Der Einsatz von hybriden Ansätzen ist noch sehr selten (Tsytsarau & Palpanas, 2012; Medhat et al., 2014).

Abschließend sei noch die Klassifikationsleistung der Sentiment Analysis angesprochen. Die Klassifikationsleistung wird meist angegeben, als Prozentanteil der Dokumente eines annotierten Test-Korpus, der korrekt klassifiziert wird (Tsytsarau & Palpanas, 2012). Weitere Metriken und das Vorgehen bei SA-Evaluationen werden in Kapitel XX näher erläutert. Grundsätzlich gilt für die Klassifikationsleistung (oft accuracy, Genauigkeit genannt), dass diese mindestens über der Leistung einer zufallsbasierten Klassifizierung liegen sollte, also bei einer binären Zuweisung (positiv vs negativ) über 50%. Als weitere Benchmark schlägt Ogneva (2012) den Übereinstimmungsgrad menschlicher Bewerter vor. Die grundsätzliche Idee ist dabei, dass ein SA-System nur so gut sein kann, wie menschliche Bewerter bezüglich eines Dokuments übereinstimmen (Ogneva, 2012). Für ihren Anwendungsfall stellt sie fest, dass menschliche Bewerter im Schnitt bei 79% der Dokumente in der Klassifikation übereinstimmen. Kim und Hovy (2004), Wilson, Wiebe und Hoffmann (2005) und Marshall (2009) berichten für unterschiedliche SA-Ebenen und -Gebieten von ähnlichen Ergebnissen um die 80%. Ein konkreter Vergleich von SA-Ansätzen auf Basis der Klassifikationsleistung ist aufgrund der unterschiedlichen Datensätze und Evaluationsframeworks sehr schwer und oft nicht möglich (Tsytsarau & Palpanas, 2012). Ravi (2015) und Tsytsarau und Palpanas (2012) berichten von Klassifikationsergebnissen zwischen 65 und 90%. Innerhalb der Lernalgorithmen erlangen SVMs im Schnitt die besten Ergebnisse (Tsytsarau & Palpanas, 2012; Chauhan Ashish & Patel, 2015). Wenn SA-Lexika mit ML-Algorithmen verglichen werden, weisen die ML-Algorithmen meist die deutlich bessere Leistung auf (Giendl & Liegl, 2008), weswegen die Methodik empfohlen wird, wenn ausreichend große Trainingsdaten vorliegen und die SA nur auf einer spezifischen Domäne ausgeführt werden soll.

Neben der Entwicklung und Analyse von SA-Algorithmen wurde auch komplexere kommerzielle und nicht-kommerzielle Anwendungen implementiert um mit Hilfe von SA verschiedene Dienste anzubieten. Vinodhini und Chandrasekaran (2012) identifizieren einige konkrete Anwendungen in der Forschung: Ku, Liang und Chen (2006) untersuchen News- und Blog-Artikel mittels SA um Meinungszusammenfassungen zu einem Thema herzustellen. Die Meinungen werden als repräsentative Sätze formuliert und mittels einer Zustimmungskurve entlang einer Achse Zu- und Ablehnung illustriert. Li und Wu (2010) integrieren SA in ihren Text-Mining zur Identifikation von Online-Foren-Hotspots. Ein weiteres Anwendungsgebiet für das SA eingesetzt wird ist Online-Advertising (Qui et al., 2010). Xu et al. (2011) nutzen SA auf Produkt-Reviews um vergleichende Beziehungen zwischen konkurrierenden Produkt-Herstellern zu extrahieren und zu visulaisieren. Damit soll die Marktanalyse und das Risiko-Management eines Unternehmens unterstützt. Red Opal ist ein Tool mit dem Nutzer Produkte im Netz nach Merkmalen suchen können. Das Tool identifiziert Produkt-Merkmale und bewertet diese gemäß der Analyse von Produkt-Reviews (Scaffidi et al., 2007). Mit dem Prototyp Opinion Observer (Liu, Hu & Cheng, 2005) können Nutzer die zusammengefassten Bewertungen von verschiedenen Produkten miteinander vergleichen. Hersteller können einen Eindruck über die Wahrnehmung eigener Produkte gewinnen. Bei Abbasi, Hassan und Dhar (2014) findet man eine Übersicht über SA-Tools für Twitter um Trends und Meinungen zu analysieren. Ein frei verfügbares SA-Tool für kurze Texte ist SentiStrength. Ferner findet man SA-Lösungen kostenfrei auch integriert in das Natural Language Toolkit (NLTK) für Python. Eine Übersicht und Evaluation von kommerziellen SA-Lösungen findet man bei Cieliebak, Dürr und Uzdilli (2013).

## Lexikon-basierte Ansätze

Im vorliegenden Projekt ist die Verwendung von Sentiment-Lexika die zentrale verwendete Methodik für die Sentiment Analysis. Aus diesem Grund werden im folgenden die wichtigsten Grundlagen des Lexikon-basierten Ansatzes erläutert, die wichtigsten Herangehensweisen zur Genese von SA-Lexika beschrieben, Vor- und Nachteile aufgezeigt, einige bekannte englischsprachige SA-Lexika beschrieben und Evaluationsergebnisse besprochen. Im nächsten Kapitel werden die wichtigsten deutschsprachigen SA-Lexika, von denen die meisten auch in der vorliegenden Arbeit verwendet werden, erläutert.

Der Lexikon-basierte Ansatz ist neben den ML-Ansatz die wichtigste zweite große Methodik zur Herangehensweise bei der Sentiment Analysis. Musto, Semeraro und Polignano (2014) nennen den Lexikon-basierten Ansatz auch den unüberwachten Ansatz in Abgrenzung zu den überwachten Lernverfahren, da keine annotierten Trainingsdaten notwendig sind. Die theoretische Grundlage des Lexikon-basierten Ansatzes geht davon aus, dass man das Sentiment eines Dokumentes, Satzes oder einer anderen Ebene auf Basis der Sentiment-Ausrichtung der einzelnen Wörter oder Phrasen bestimmen kann, die das Dokument oder den Satz konstituieren (Taboeda et al., 2011; Musto et al., 2014). Als Sentiment-Ausrichtung wird im Folgenden die Polarität, also die Ausprägung bezüglich positiver, negativer oder neutraler Ausrichtung verstanden, da dies die vorherschende Verwendungsweise in der SA ist (siehe Kapitel XX). Wörter und Phrasen nun die eine positive oder negative Polarität hervorrufen werden Sentiment-Wörter oder auch Sentiment-Tragende Wörter genannt (Liu, 2016, S. 189). Im Folgenden wird der englische Begriff Sentiment-Bearings Words (SBWs) genutzt. Beispiele für positive SBWs wären gut, wundervoll oder schön; für negative SBWs schlecht, furchtbar oder hässlich. Liu (2016, S. 189) nennt als Beispiel für eine englischsprachige Sentiment-tragende Phrase *cost an arm and leg.* Ähnliche Sprichwörter sind auch für das Deutsche denkbar. Oft werden SBWs auch als Wörter definiert die einen wünschenswerten Zustand (positiv) oder einen unerwünschten Zustand (negativ) ausdrücken (Ding et al., 2008, Liu, 2016, S. 189). Diese Definition greift besonders bei der Betrachtung von Reviews.

Eine Liste derartiger Wörter und Phrasen wird als Setiment-Lexicon bezeichnet (Liu, 2016, S. 10). Innerhalb des Lexikons wird für jedes Wort die sogenannte Prior-Polarität angegeben, also die kontextunabhänge Polaritätsausrichtung des Wortes (Wilson et al., 2005). Diese Ausrichtung kann auf unterschiedliche Weise angegeben werden, z.B. als numerischer Wert bezüglich der Polaritätsklassen auf einer metrischen Skala mittels eines sogenannten Sentiment-Scores (SentiWordNet; Esuli & Sebastiani, 2007; Baccianella, Esuli, & Sebastiani, 2010) oder als dichotome Zugehörigkeit zu einer Polaritätsklasse ohne Angabe der Intensität (NRC, Mohammad & Turney, 2013). Als weiteres Beispiel für eine Angabemöglichkeit der Priori-Polarität sei das MPQA-Lexikon genann (Wilson, Wiebe & Hoffman, 2005). Hier wird die Polarität eines Wortes (positiv, negativ oder neutral) und die Intensität über die dichotome Skala stark (strong) und schwach (weak) angegeben. Es gibt keine feste Standardisierung (Emersen & Declerck, 2014).

Die Gesamtpolarität eines Dokumentes, eines Satzes oder eines anderen Untersuchungsgegenstandes der SA wird über beim Lexikon-basierten Ansatz nun über die Nutzung eines (oder mehrerer) SA-Lexika berechnet. Die Gesamtpolarität kann dabei als Summe der Polaritätswerte der einzelnen SBWs berechnet werden (Palanisamy, Yadav & Elchuri, 2013). Bei der SA auf Satzebene wird bei der dichotomen Angabe der Polarität ein Satz beispielsweise als positiv angenommen, der fünf positive Wörter enthält und zwei negative oder bei Angabe von Polaritäts-Intensitäten, dessen Summe an Negativ-Scores, die der Positiv-Scores übersteigt. Die Gesamt-Scores eines Dokumentes oder Satzes können dann als Polaritätsausprägung auf einer stetigen Skala angegeben werden (Kennedy und Inkpen, 2006; Nachweis?, Tsytsaru). Im erstgenannten Beispiel hätte der Satz eine negative Ausprägung von fünt, eine positive Ausprägung von zwei und gemäß Summenbildung eine negative Gesamtausprägung von drei bzw. –3. Eine derartige Methode, die ein Sentiment-Lexikon ohne Intensitäten verwendet wird auch häufig als Term-zählende-Methode (Term-Counting-Method) bezeichnet, da lediglich die Summe an positiven und negativen Wörtern Einfluss auf das Endergebnis hat (Turney, 2002; Kennedy und Inkpen, 2006). Ein Dokument mit mehr positiven Wörtern wird demnach als positiv angenommen, ein Dokument mit mehr negativen Wörtern als negativ (Tuney, 2002). Werden Polaritäts-Intensitäten zur Summenbildung genutzt, kann ein Dokument auch negativ sein wenn es weniger negative Wörter hat diese aber insgesat ein stärkeres Gewicht haben als die häufigereren positiven Wörter insgesamt.

Um die Gesamt-Scores von Dokumenten oder Sätzen vergleichbar zu halten, können die Werte an der Zahl von Untereinheiten, z.B. der Zahl von Wörtern normalisiert werden (Musto et al., 2014). Somit wird ein Durchschnittswert gebildet der gemäß Normalisierung verschiedene Dokumente oder Sätze vergleichbar macht und eine tiefergehende Wert-Analyse erlaubt. Zu Beginn der Forschung wurden vor allem Adjektive als zentrale SBWs betrachtet, mittlerweile werden jedoch fast alle Wortformen als potentielle SBWs betrachtet, vor allem noch Nomen, Verben und Adverben (Taboada et al., 2011).

Neben der reinen Ad-Hoc-Verwendung eines Lexikons und den oben beschriebenen Kalkulationsansätzen findet man in der Forschung einige Methoden zur Verbesserung des Ansatzes. Über regelbasierte Verfahren wie in Kapitel XX geschildert kann man den Einsatz von Sentiment-Lexika verfeinern und präzisieren. Dazu gehört die Identifikation von Wörtern im Umfeld eines SBWs, die die Polarität beeinflussen und die Integration dieser Wörter in die Polaritätsberechnung. Kennedy und Inkpen (2006) benennen die Gruppe dieser Wörter als Valence-Shifters (VS). Zu den VS-Wörtern zählen sie Negationen, Verstärker (Intensifier) und Verminderer (Diminisher). Intensifier werden manchmal auch Amplifier genannt und Diminisher Downtoners (Taboada et al., 2011). Negationen drehen die Polaritätsausprägung um, ein positives Word ist dann negativ z.B. beim Ausdruck „nicht schön“. Intensifier verstärken die Polaritätswirkung, z.B. „sehr schön“ und Diminisher verringern diese, z.B. „kaum schön“. Implizite Valence-Shifter eines Satzes sind Ironie und Sarkasmus. Einige Projekte untersuchen Möglichkeiten der Identifikation von Ironie auf linguistischer Ebene (Reyes, Rosso & Veale 2013; Nachweis) und den Umgang mit dieser in der SA (Polanyi & Zaenen, 2006; Nachweis). Aufgrund der anspruchsvollen Aufgabe wird Ironie beim Einsatz von SA-Lexika jedoch selten beachtet (Kennedy & Inkpen, 2006).

Eine andere Methode, die bei der Nutzung von SA-Lexika zur Optimierung verwendet wird, ist die Lemmatisierung des zu analysierenden Textes (Nachweis). Dies ist beispielsweise ein notwendiger Schritt, wenn ein Lexikon nur die Grundformen von Wörtern enthält, so dass flektierte Formen eines SBWs auch gefunden werden (Kennedy & Inkpen, 2006). Mehr zur Lemmatisierung in der SA wird in Kapitel XX besprochen. Ein anderer häufiger Schritt der Vorverarbeitung ist das Part-Of-Speech-Tagging. Dabei wird die Wortart jedes Wortes bestimmt. Dies kann so genutzt werden, dass nur eine bestimmte Wortart für die SA genutzt wird oder das unterschiedliche Wortarten unterschiedlich gewichtet werden (Nachweis). Beim Einsatz von Lexikon-basierten Verfahren kann die Vorverarbeitung des zu analysierenden Textes als Verarbeitungskette mit mehreren Schritten betrachtet werden (Nachweis + Beispiel). Fortgeschrittene Ansätze versuchen des Weiteren den Lexikon-basierten Ansatz durch Kombination mehrere Lexika (Taboada et al., 2011; Emersen & Declerck, 2014) oder durch Synonymerweiterung zu verbessern (Kennedy & Inkpen, 2006; Agarwal et al., 2011).

Zum Verständnis der Funktionalität sowie der Vor- und Nachteile von Sentiment-Lexika werden nachfolgend die Grundlagen der Lexikon-Kreation erläutert. Liu (2016, S. 189-201) identifiziert drei Hauptherangehensweisen: den manuellen Ansatz, der Wörterbuch-basierte Ansatz (dictionary-based approach) und der Korpus-basierte Ansatz (corpus-based approach). Bei einem manuellen Ansatz werden die Wörter und Polaritäten für das Lexikon manuell gesammelt. Dieser Ansatz ist sehr zeit- und arbeitsaufwendig, meist werden manuelle Arbeitswesen nur zur Überprüfung automatisierter Verfahren genutzt (Liu, 2016, S. 189; Nachweis). Mohammad und Turney (2013) nutzen Crowdsourcing über Amazon Mechanical Turk zur manuellen Sentiment-Bestimmung in ihrem Prozess zur Kreation eines Sentiment-Lexikons.

Die grundsätzliche Technik beim Wörterbuch-Ansatz ist es einige bekannte und klare Sentiment-Wörter wie gut und schlecht als Ausgangswörter zu sammeln und dann das Lexikon so aufzubauen, indem man in Online-Wörterbüchern nach Synonymen und Antonymen sucht. Das Verfahren kann dann iterativ weiter geführt werden indem wiederum nach den Synonymen und Antonymen der neu gefundenen Wörter gesucht wird bis zu einem Abbruch (z.B. wenn keine weiteren Wörter mehr gefunden werden; nach Liu, 2016, S. 190). Durch manuelle Inspektion können die Ergebnisse noch auf Fehler untersucht werden. Verschiedene fortgeschrittene Methoden ermöglichen es über diesen Ansatz auch Sentiment-Intensitäten, also den Grad der Polarität zu bestimmen, beispielsweise über eine probabilistische Methode (Kim & Hovey, 2004) oder Distanz- und Graph-Algorithmen zwischen den Ausgangswörtern und den gefundenen Wörtern (Kamps et al., 2004; Blair-Goldensohn et al., 2008). Lexika, die auf diese Weise erstellt werden, werden meist General-Purpose-Lexicon genannt und sollen und können vom Prinzip her für alle Anwendungsgebiete verwendet werden (Liu, 2016, S. 191). Dazu im Gegensatz wird über den Korpus-basierten Ansatz versucht Sentiment-Lexika für eine spezifische Domäne zu generieren, also Lexika, die die kontext-spezifischen Sentiment-Besonderheiten von Wörtern einer bestimmten Domäne beachten. Dabei stellt Liu (2016, S. 191-197) zwei Hauptszenarios fest: Weitere domänenspezische Sentiment-Wörter zu finden und kontext- und domänenspezifische Änderungen von Sentiment-Wörtern aus General-Purpose-Lexika zu identifizieren. Hierfür gibt es verschiedene Herangehensweisen. Hatzivassiloglou und McKeown (1997) nutzen beispielsweise linguistische Regeln in einem Domänenkorpus um weitere Sentiment-Wörter zu finden. Sie starten wieder mit einer Sammlung bekannter Ausgangswörter und sammeln Wörter unter anderem über den Ansatz dass mit „und“ verknüpfte Wörter mit den Ausgangswörtern meist dieselbe Polarität haben. Ein anderer Ansatz nutzt syntaktische Relationen von Meinungen und Meinungszielen in domänenspezifischen Texten aus um SBWs zu identifizieren (Qiu et al., 2009; 2011). Ding et al. (2009) versuchen den domänenspezifischen Kontext in die SA zu integrieren indem statt das einzelne SBW ein Tupel aus SBW und Bezugsaspekt gebildet wird. So kann ein und dasselbe SBW mit unterschiedlichen Bezugsaspekten unterschiedliche Polaritäten besitzen. Weitere Herangehensweisen, Beispiele und ausführliche Beschreibungen der genannten Methonden zur Sentiment-Lexikon-Genese findet man bei Liu (2016, S. 189-201) und bei Medhat et al. (2014). Fortgeschrittene hybride Ansätze oder Kombinationen mit regelbasierten Verfahren wurden bereits in Kapitel XX angesprochen.

Zahlreiche englischsprachige SA-Lexika wurden mit ähnlichen Methoden entwickelt und werden in der Forschung zur SA eingesetzt. Einige der wichtigsten werden nachfolgend genannt. Das General Inquirer lexicon (Stone, 1968) ist eines der ältesten für die SA genutzte Lexikon. Es handelt sich um ein manuell erstelltes Lexikon, dass auf Arbeiten der Kognitionspsychologie von Wortbedeutungen und Inhaltsanalyse basiert. Es basiert auf manuellen Ratings von Studien-Teilnehmern. Das General Inquirer lexicon enthält zu jedem Wort die dazugehörigen semantischen Kategorien, von denen einige, vor allem die Tags positiv und negativ für die SA relevant sind. Jurafsky und Martin (2016) zählen 1915 positive Terme und 2291 negative. Das Tool ist frei verfügbar. Esuli und Sebastiani bezeichnen das Lexikon 2007 als die Benchmark in der Term-basierten Sentiment-Klassifikation und als das größte manuell annotierte Lexikon. Das MPQA Subjectivity lexicon (Wilson et al., 2005) enthält 2718 positive and 4912 negative. Es wurde durch über mehrere Methdoen erstellt, unter anderem die Kombination bestehender Datensätze und die manulle Auszeichnung. Es enthält für jedes Wort die Wortart, die Polarität (positiv, negativ, neutral) und die Intensität in der Einteilung strong (stark) und weak (schwach). Die Ressource SentiWordNet (Esuli & Sebastiani, 2007) gibt für jedes Synset des Online-Lexikons WordNet einen numerischen Wert für die Objektivität (neutral), die Positivität und Negativität von 0 bis 1 an. Ein Synset in WordNet ist eine Gruppe von Synonymen, die das gleiche Konzept repräsentieren. In Synset kann ein Term zu mehreren Synsets gehören, was in SentiWordNet zu Ambiguitäts-Problemen führt (Musto et al., 2014). WordNet besteht aus etwa 115 000 Synsets, SentiWordNet enthält die obigen Polaritätsangaben zu allen Synsets. Das Lexikon ist frei verfügbar und Tsytsarau und Palpanes (2011) bezeichnen SentiWordNet als das zur Zeit populärste Sentiment-Lexikon in der SA.

Als weitere wichtige Beispiele seien noch das auf Konsumenten-Reviews basierte Sentiment-Lexicon von Hu und Liu (2004), das kostenpflichtige Linguistic Inquiry and Word Counts (Pennebaker et al., 2007), das für Microblogs angepasste AFFIN (Nielsen, 2011), die lexikalische Ressource SenticNet (Cambria et al., 2010) und das NRC-Emotion-Lexicon (Mohammad & Turney, 2013) genannt. Die beiden letztgenannten zeichnen sich dadurch aus, dass sie neben der Polarität Wörter auch auf komplexe emotionale Kategorien (z.B. Zorn, Traurigkeit, Freude) abbilden. Eine Beschreibung der genannten als auch weiterer englischsprachiger Lexika findet man bei Musto et al. (2014), D‘ Andrea et al. (2015), Liu (2016, S. 200-201) und Jurafsky und Martin (2016, S. XX). Die in der vorliegenden Arbeit verwendeten deutschsprachigen Lexika werden im nachfolgenden Abschnitt besprochen.

Die Nutzung von SA-Lexika weist einige Vor- und Nachteile auf. Im Vergleich zum ML-Ansätzen ist der Vorteil, dass keine Trainingsdaten gesammelt werden müssen und die Ad-Hoc-Verwendung generell eher einfach ist und schnell geht (D‘ Andrea et al., 2015). Der Nachteil ist, dass man lediglich eine endliche Liste von Wörtern mit fixen Polaritätswerten nutzt, die kontextabhängige Ambiguitäten und Besonderheiten nicht beachten (D‘ Andrea et al., 2015). Dies ist vor allem bei General-Purpose-Lexika der Fall. Domänenspezifische Lexika können diesen Umstand etwas ausgleichen, sind jedoch nicht auf außerhalb der Domäne liegende Anwendungsfälle anwendbar. Des Weiteren ist die Funktionalität der SA-Lexika auch von der korrekten Vorbereitung des zu analysierenden Textes über z.B. Lemmatisierung abhängig. Auch die Genese von SA-Lexika muss bei der Bewertung der Methodik beachtet werden. Als Benchmark für die Gültigkeit wird die manuelle Annotation von SBWs durch Menschen angesehen (Nachweis), die Mehrzahl der aktuelleren SA-Lexika wurde jedoch maschinell erstellt und nicht manuell korrigert, was zu Ungenauigkeiten und Fehlern führen kann. Liu (2016, S. 10-11) beschreibt noch weitere Probleme anhand von konkreten Beispielen: Ein und dasselbe Wort kann in verschiedenen Domänen und Kontexten völlig unterschiedliche Bedeutungen haben, sogar Polaritätswechsel aufweisen. Auch ist nicht jeder Satz der SBWs enthält zwingen Ausdruck einer Meinung oder eines Gefühls. Als drittes großes Problem identifiziert Liu Sarkasmus. Dies wurde bereits weiter oben angesprochen. Ferner gibt Liu (2016, S. 11) Beispiele für Sätze an, die keine SBWs enthalten jedoch Meinungen und Gefühle ausdrücken. Dies kann beispielsweise durch den Einsatz von Methaphern geschehen. Spezielle Probleme von SA-Lexika für die Ziele der vorliegenden Arbeit werden in Kapitel XX genauer erläutert.

Bereits in Kapitel XX wurden Evaluationsergebnisse beim Vergleich von ML- und Lexikon-Ansätzen erläutert. Eine Zusammenfassung von Evaluationsergebnissen in der Forschung zu unterschiedlichen SA-Ansätzen findet man Tsytsarau und Palpanas (2011). Für Lexikon-Ansätze identifizieren sie, je nach Anwendungsgebiet und konkreter Methodik Ergebnisse zwischen ca. 40% und 87% Prädiktionsgenauigkeit. Aufgrund der unterschiedlichen Evaluationsframeworks und Anwendungsgebiete sind Evaluationsergebnisse nur bedingt vergleichbar (Tsytsarau & Palpanas, 2011). Wurden in Studien ML-Algorithmen mit SA-Lexika unter den gleichen Bedinungen (gleicher Evaluationskorpus, gleiches Evaluationsframework) verglichen, schneiden ML-Algorithmen besser ab (Giendl & Liegl, 2008; Pang et al., 2002). Als konkretes Beispiele mit guten Ergennungsraten seien hier noch Godbole et al. (2007) mit Genauigkeitsraten zwischen 82% - 96% bei der SA von Blog-Posts mittels eines erweiterten Wörterbuch-Ansatzes genannt. Auf Basis der Methodenanalyse der SA-Literatur und -Forschung ist die Verwendung von SA-Lexika trotz aller Kritik die geläufige SA-Methode bei nicht vorliegenden Trainingsdaten.

Einige Studien verfolgen die Evaluationsidee ein eigens erstellte SA-Lexikon mit bekannten anderen Lexika oder einen entwickelten Lexikon-basierten Ansatz mit unterschiedlichen SA-Lexika zu vergleichen. Musto et al. (2014) vergleichen verschiedene Lexika bezüglich ihres Normalisierungsansatzes und stellen fest, dass SentiWordNet die beste Erkennungsgenauigkeit (ca. 58-59%) auf einem annotierten SemEval2013-Korpus (Nakov et al., 2013) aufweist. Aud dem ebenfalls mit Sentiment-Informationen ausgezeichneten Stanford-Twitter-Sentiment-Datensatz (Go, Bhayani & Huang, 2009) ist das Lexikon SenticNet mit einer Genauigkeit von bis zu 75% das beste SA-Lexikon. Khoo und Jonkhan (2017) vergleichen das eigene General-Purpose-Lexikon mit anderen gängigen Lexika. Die Lexika sind erbringen in etwa die gleiche Leistung auf einem Produkt-Review-Korpus (75-77% Genauigkeit). Für ein News-Headline-Korpus erzielt das eigene Lexikon das beste Ergebnis (69%). Ferner sei noch Nielsen (2011) genannt der seine eigene Wortliste mit anderen vergleicht. Er nutzt einen mit Intensitätsangaben annotierten Test-Korpus mit Stärke-Werten zwischen 1 und 9 (Biever, 2010) und kann deswegen Korrelationskoeffizienten statt Genauigkeitsangaben nutzen. Das unter anderem auf einer Wortliste basierende Tool SentiStrength (Thelwall et al., 2010) erbringt dabei die beste Leistung. Die Idee des systematischen Lexika-Vergleichs wird in der vorliegenden Arbeit aufgegriffen. Ein derartiger Vergleich deutschsprachiger SA-Lexika wie im nächsten Kapitel beschrieben, ist nicht bekannt.

#Evaluations-Datensätze

## Deutschsprachige Sentiment-Lexika

In diesem Abschnit werden die wichtigsten in der Forschung identifizierten Sentiment-Lexika zum Einsatz in deutschsprachigen Anwendungsfeldern beschrieben. Die Verfügbarkeit nicht englischsprachiger Sentiment-Lexika ist im Vergleich sehr gering (Nachweis). Für jedes Lexikon wird die Erstellung, der grundlegende Aufbau und der Einsatz in der Forschung erläutert. In einem abschließenden Kapitel werden Auffälligkeiten und Erkenntnisse zusammengefasst. Ein Großteil der beschriebenen Lexika wird in der vorliegenden Arbeit verwendet und beispielsweise in Kapitel XX noch genauer mit textanalytischen Methoden analysiert.

#Zahlenangaben werden später konkretisiert

### SentimentWortschatz – SentiWS

SentimentWortschat (SentiWS) ist ein frei verfügbares deuschsprachiges Sentiment-Lexikon von Remus, Quasthoff und Heyer (2010). Das Lexikon enthält 1650 negativ und 1818 positiv annotierte Sentiment-tragende Wörter, also insgesamt 3468 Wörter. Das Lexikon ist derart aufgebaut, dass zunächst jedes Wort mit seiner Grundform aufgelistet ist. Danach wird die Wortart als POS-Tag nach dem Stuttgart-Tübingen-Tagset (Thielen et al., 1999) angegeben sowie die Polarität als numerischer, metrischer Wert zwischen -1 (maximal negativ) und +1 (maximal positiv). Es enthält Adjektive, Adverbien, Nomen und Verben. Des Weiteren werden für einige Wörter die flektierten Wortformen der Grundform angegeben, also z.B. das Präteritum bei Verben. Diese wurden über eine interne Datenbank bezogen. Remus et al. (2010) geben jedoch an, dass die Korrektheit und Vollständigkeit der Flektionsformen nicht garantiert wird. Aufgerechnet mit den flektierten Formen enthält das Lexikon gemäß Remus et al. (2010) 16 406 positive und 16 328 negative Wörter, was insgesamt 32 734 Wörter ergibt. Folgender Screenshot illustriert den Aufbau des Lexikons:

Das Lexikon wurde auf Basis von drei Quellen generiert. Zunächst wurden die positiven und negativen Kategorien des bereits angesprochenen (siehe Kapitel XX) General Inquirer lexicon akquiriert und semi-automatisch mittels Google Translate und einer anschließenden manuellen Überprüfung übersetzt. Ferner wurden manuell einige spezielle Wörter aus dem Finanzbereich hinzugefügt, da dies die ursprüngliche Zieldomäne des Lexikons ist. Die zweite Wort-Quelle basiert auf der textuellen Analyse von als negativ und positiv ausgezeichneten Produkt-Reviews. Über textanalytische Methoden wurden die Wörter gesammelt, die in besonders vielen positiven bzw. negativen Reviews vorkamen und dann manuell, bei entsprechender Eignung, dem SBW-Wortschatz hinzugefügt. Als letzte Quelle wurde das German Collocation Dictionary (Quasthoff, 2010) genutzt. Dieses gruppiert deutschsprachige Wörter bezüglich ihrer semantischen Nähe. Über die bislang aus den anderen beiden Quellen akquirierten Wörter konnten so weitere semantisch nahe Wörter gefunden wurden, die final auch noch mal manuell auf ihre Eignung als SBW überprüft wurden. Die einzelnen Polaritätsgewichte wurden sodann mittels der semantischen Nähe zu Ausgangswörter (Seed-Words) berechnet. Es wurden einige besonders starke, eindeutige positive und negative Ausgangswörter definiert und dann über die statistische Analyse eines großen deutschsprachigen Korpus festgestellt wie häufig Wörter des Lexikons in der Nähe dieser Seed-Words erscheinen. Grundsätzlich gilt also je häufiger ein Wort in Sätzen mit den Seed-Words erscheint desto höher ist die jeweilige Polarität. Das Verfahren im Detail wird bei Remus et al. (2010) beschrieben. Ein sehr positives Wort ist beispielsweise Freude (+0,6502), ein sehr negatives schädlich (-0,9269).

Zur Evaluation des erstellten SA-Lexikons wurde manuell ein Korpus aus 480 Sätzen aus Internet-Foren erstellt. Zwei menschlichen Beurteilern haben sodann alle Wörter in den Sätzen bezüglich ihrer Polarität (positiv vs negativ vs neutral) beurteilt. Die Annotationen der Bewerter wurden dann gegen die Angaben in SentiWS geprüft. Auf diese konnten folgende Gesamt-Evaluationsergebnisse konstatiert werden: Präzision = 0,96, Recall = 0,74, F-Wert = 0,84 (Eine Erklärung der Metriken findet man in Kapitel XX). Insgesamt erbringt SentiWS also eine sehr gute Leistun gemäß dem gewählten Evaluationsverfahren. Es erzielt bessere Ergebnisse bei der Betrachtung von nur negativen Wörtern. Problematisch für die Leistung sind für SentiWS domänenspezifische Begriffe, nicht-deutsche Begriffe, orthographische Fehler und Besonderheiten und ambige Wörter.

#BeispielForschung 😊

### Berlin Affective Word List – Reloaded (BAWL-R)

Die Berlin Affective Word List (BAWL) ist ein frei verfügbarer Datensatz aus der psychologischen Forschung, der zur Sentiment Analysis genutzt werden kann. Die erste Version des Bawl (Vo, Jacobs & Conrad, 2006) wurde von Vo et al. (2009) erweitert und optimiert und ist als Berlin Affective Word List – Reloaded (BAWL-R) bekannt. In der vorliegenden Arbeit wird das BAWL-R verwendet und es wird im weiteren Bawl genannt.

Der Datensatz besteht aus einer tabellarischen Liste von Wörtern. Jedes Wort ist einmal mit Großbuchstaben und Kleinbuchstaben angegeben. Die Wortart ist über die Buchstaben N für Nomen, V für Verben und A für Adjektive angegeben. Zu jedem Wort sind nun zahlreiche Werte basierend auf Konzepten der Psychologie angegeben. Zentral für die Sentiment Analysis sind die Angaben gemäß der Valenz-Arousel-Theorie (Nachweis). Es handelt sich um zwei orthogonale Dimensionen zur Bewertung von Emotionen mit ruhig bis erregt für Arousel und positiv/angenehm bis negativ/unangenehm für die emotionale Valenz. Die durchschnittliche Bewertung der Wörter gemäß dem noch weiter unten beschriebenen Bewertungsverfahren gemäß der genannten Kategorien können für die Sentiment-Analysis verwendet werden. Die Valenz kann von -3 bis 2,9 verlaufen, Arousel von 1,1 – 4,7. Die sonstigen Angaben im BAWL-R werden nicht weiter genutzt. Bezüglich der Polarität wird die Valenz als äquivalent betrachtet, Werte unter 0 stehen dabei für negative Polarität, Werte über 0 für positive Polarität. Insgesamt sind 2902 Wörter enthalten. Eigene Analysen (siehe Kapitel XX) ergaben, dass es insgesamt 2107 Nomen, 502 Verben, 291 Adjektive enthält. Davon fallen 1576 in die positive und 1266 in die negative Kategorie. Ferner gibt es 60 Wörter die einen Valenz-Wert von 0 haben. Ein besonders negatives Wort ist Krieg (-2,9), ein sehr positives Liebe (2,9), ein Beispiel für ein neutrales ist Maus (0). Ein Beispiel für ein Wort mit sehr hohem Arousel ist Attentat (4,7); das Wort still (1,2) hat ein sehr geringes Arousel. Folgender Screenshot zeigt das Format des Bawl auf. EMO\_MEAN steht dabei für die Valenz und AROUSEL\_MEAN für Arousel.

Die Wörter des ursprünglichen BAWL (Vo et al., 2006) wurden aus der CELEX-Datenbank bezogen (Baayen, Piepenbrock, & van Rijn, 1993). Das ist eine Online-Datenbank, die Wörter mit linguistischen Informationen enthält. Die Wörter wurden von studentischen Teilnehmern bezüglich ihrer emotionalen Valenz und anderer Konzepte in einer Studie bewertet. Die Skala für emotionale Valenz wurde auf einer ordinalen Skala von -3 bis 3 angegeben. Die Mittelwerte ergeben somit die Polarität eines Wortes. Im erweiterten Sinne handelt es sich also beim Bawl um ein manuell erstelltes Lexikon, mit durchschnittlichen Annotationen von Studienteilnehmern. Vo et al. (2006) nutzen Teile des BAWL um ein psychologisches Experiment durchzuführen. Vo et al. (2009) erweitern und optimieren die erste Version und präsentieren das BAWL-R. Über den gleichen Ansatz wie zuvor wurde zunächst die Wortanzahl erhöht und weitere Bewertungen bezüglich psychologischer Konzepte gesammelt. 200 Teilnehmer beurteilten die neuen Wörter wieder, analog zu oben, bezüglich der emotionalen Valenz aber nun sowohl die alten als auch die neuen Wörter bezüglich Arousel auf einer 5-stufigen Skala von geringem Arousel bis hohem Arousel. Da das deutsche Wort Erregung für Arousel ungeeignet ist wurde SAM (Lang, 1980) verwendet. Dabei handelt es sich um ein Strichmännchen, über das man bildhaft steigende Arousel-Zustände angeben kann. Die jeweiligen Mittelwerte stellen wieder die für ein Wort annotierten Werte für die emotionale Valenz (hier mit der Polarität gleichgesetzt) und Arousel dar. Weitere Daten für psychologische Konzepte sowie linguistische Informotionen wurden gesammelt und sind Bestandteil des Lexikons, spielen jedoch für die weitere SA keine Rolle.

#Forschung bezüglich Bawl

### NRC Word-Emotion Association Lexicon (NRC)

Das NRC Word-Emotion Association Lexicon (NRC), auch EmoLex, genannt ist ein im Original enlischsprachiges Sentiment-Lexikon von Mohammad und Turney (2010; 2013a, 2013b). Es besteht aus 14 182 Wörtern die bezüglich Polarität, also positiv oder negativ, sowie acht Basis-Emotionskategorien annotiert sind. Die Wörter werden von Mohammad und Turney (2010) Target-Words genannt, Polarität und Emotionskategorien Affect-Category. Das NRC gibt für ein Wort an, ob es mit einer Affekt-Kategorie assoziiert wird. Die Annotation liegen dichotom vor, mit den Werten 0 für „keine Assoziation“ und 1 für „Assoziation vorhanden“. Es werden keine Intensitäten angegeben. Wörter können 0 – 10 Assoziationen haben, also gar keine Assoziation haben oder auch mehrere. Eigene Analysen ergaben, dass Wörter sowohl positiv als auch negativ assoziert sein können. Die acht Emotionskategorien basieren auf der Emotionstheorie von Plutchik (1980). Er schlägt die Emotionskategorien Wut (anger), Furcht (fear), Freude (joy), Traurigkeit (sadness), Ekel (disgust), Überraschung (surprise), Vertrauen (trust) und Erwartung (anticipation). Die englischen Original-Begriffe für die Emotionskategorien wurden nach eigenem Ermessen übersetzt. Das Lexikon, sowie weiter dazugehörige Datensätze, sind frei über Anfrage verfügbar.

Es wurden fremdsprachige Versionen mittels einer automatischen Übersetzung über Google Translate erstellt. Die deutschsprachige Version wird in der vorliegenden Arbeit verwendet. Die Größe des Lexikons reduziert sich nach Angaben der NRC-Webseite auf 11812 Einträge. Es ist nicht eindeutig klar, wodurch diese Reduktion entsteht. Es handelt sich jedoch möglicherweise um Phrasen und Wörter die nicht automatisch übersetzt werden konnten. Spätere Analysen (siehe Kapitel XX) zeigen, dass durch die Übersetzung doppelte Wörter entstehen. Insgesamt hat das deutsch NRC folglich 9629 unterscheidbare Terme. Folgende Tabelle gibt für die verbliebenen Wörter die Termverteilung pro Affekt-Kategorie an. Terme könne dabei mit mehreren aber auch keinen Affekt-Kategorien assoziert sein.

|  |  |  |
| --- | --- | --- |
| Affekt-Kategorie | Term-Häufigkeit | Beispiel |
| Positiv | 1706 | himmlisch |
| Negativ | 2322 | Katastrophe |
| Wut | 887 | Mörder |
| Erwartung | 653 | Wettbewerb |
| Ekel | 776 | Verstümmelung |
| Furcht | 1072 | Dunkelheit |
| Freude | 534 | jubeln |
| Traurigkeit | 868 | Tragödie |
| Überraschung | 402 | Verrat |
| Vertrauen | 977 | schwören |

Des Weiteren sind 4855 Wörter entthalten, die für alle Kategorien den Wert 0 angegeben haben, also vollkommen neutral sind. Ein Beispiel ist das Wort „Avocado“. Abzüglich einiger weiterer Sonderfälle (siehe Kapitel XX) ergibt dies 4529 Wörter, die bezüglich einer Affekt-Kategorie mindestens eine Assoziation aufweisen. 48 Wörter oder Phrasen sind als sowohl positiv als auch negativ markiert. Beispiele hierfür sind die Wörter „Bücherwurm“ oder „Revolution“. Es ist zu beachten, dass aufgrund der automatischen Übersetzung das deutschsprachige Sentiment-Lexikon sehr fehlerbehaftet sein kann. Es ist jedoch das einzige deutschsprachige SA-Lexikon mit anderen Kategorien als der Polarität, dass akquiriert werden konnte und wird deswegen in der vorliegenden Arbeit dennoch verwendet.

Das Original-NRC wurde in einem mehrstufigen Prozess erstellt, der im Detail bei Mohammad und Turney (2010; 2013a, 2013b) nachgelesen werden kann. Die grundsätzliche Idee ist ein Crowdsourcing-Ansatz. Die Erstellung der Wort- und Phrasengrundlage basiert auf mehreren Ausgangsquellen und umfasst manuelle Kontroll- und Anpassungsschritte. Unter anderem wurden Wörter aus dem Macquarie Thesaurus, dem General Inquirer lexicon und dem WordNet Affect Lexicon bezogen. Über Amazon Mechanical Turk haben Teilnehmer die gewonnen Wörter bezüglich der Assoziationen zu den Affekt-Kategorien beurteilt. Teilnehmer konnten zu jedem Wort und jeder Assoziationskategorie angeben ob das Wort und die Kategorie gar nicht, schwach, moderat oder stark assoziert ist. Kontrollfragen testen die Aufmerksamkeit und die Bewertungsfähigkeit. Insgesamt konnten so Bewertungen von 2216 Teilnehmern erhoben werden. Jeder Term wurde von 5 Teilnehmern eingeschätzt. In der finalen Auswertung wurden keine oder schwache Zusammenhänge als keine Assoziation interpretiert und moderate und starke Zusammenhänge als Assoziation vorhanden. Die endgültige Annotation eines Wortes wurde dann über die Mehrheit der jeweiligen Angaben bestimmt.

# Annotator-Angaben 🡪 2010er paper

#NRC in der Forschung

### Clematide-Dictionary (CD)

Clematide und Klenner entwickelten (2010) ein deutschsprachiges, online frei verfügbares, Polaritätslexikon. Ein offizieller Name ist nicht bekannt, in der vorligenden Arbeit wird es mit CD für Clematide-Dictionary abgekürzt. Das Lexikon wurde Bi-directional Sentiment Composition der Swiss National Science Foundation (SNSF) bezogen (Klenner et al., 2010). Das Lexikon besteht im Ausgangszustand aus 8402 Termen mit Polaritätsangaben. Die Polarität wird einerseits über die Zugehörigkeit zu einer Polaritätsklasse positiv (POS), negativ (NEG) oder neutral (NEU) annotiert. Des Weiteren wird die Intensität mit zwei möglichen Belegungswerten angegeben, 1 für eine starke Assoziation und 0,7 für eine schwächere Assoziation. Es handelt sich also nicht um eine vollständige Ausreizung der metrischen Skala von 0-1 sondern eher um eine binäre Angabe, die numerisch angegeben wird. Folgender Screenshot aus der Lexikon-Datei verdeutlicht das Format der Angabe:

Das Lexikon enthält Wörter aus den Wortarten Nomen, Adjektiven und Verben sowie einige Phrasen. Insgesamt sind 2912 Einträge positiv annotiert, 4889 negativ und 594 neutral. Ferner sind noch 7 sogenannte Valenz-Shifter enthalten. Die Bedeutung und Definition von Valenz-Shiftern findet man in Kapitel XX. Es handelt sich dabei um Wörter die die Polaritätsausrichtung und -stärke verändern, wenn sie im Umfeld eines Sentiment-Wortes stehen. Es sind 5 Shifter enthalten (SHI), die die Polarität umdrehen und 2 Intensifier (INT), die die Polarität abschwächen oder stärken. Ein Beispiel für ein besonders negatives Wort ist „wehmütig“ (NEG=1), ein positives SBW ist „sympathisch“ (POS=1). Beispiele für Shifter-Wörter sind „nicht“ und „beenden“, die Intensifier sind „wenig“ und „viel“. Diese sind mittels eines Intensitätsfaktors annotiert, den man zur Multiplikation verwenden kann. Das Lexikon enthält, wie viele andere, ambig annotierte Wörter, die mehreren Polaritätsklassen zugeordnet sind. Beispiele hierfür wären die Wörter „sorgen“ und „bescheiden“, die beide sowohl positiv als auch negativ annotiert sind. Insgesamt hat das Lexikon 41 derartiger Wörter. Ferner enthält es auch 124 Phrasen, also Einträge, die aus mehr als einem Wort bestehen. Der Umgang mit diesen und anderen Sonderfällen, sowie eine tiefergehende inhaltliche Analyse findet man in Kapitel XX.

Die erste Version des Lexikons wurde durch die Akquisation und manuelle Annotation von Synsets der lexikalischen Datenbank GermaNet (Hamp & Feldweg, 1997) durchgeführt. GermaNet ist eine ähnliche Datenbank wie WordNet und die Nutzung von Synsets aus WordNet wurde bereits für englischsprachige SA in Kapitel XX angesprochen. Synsets sind semantische Konzepte zu denen mehrere Terme gehören könnn, die ein und dasselbe Konzept repräsentieren. Etwa 8000 Synsets wurden manuell bezüglich Polarität (positiv, negativ und neutral) sowie bezüglich Polaritätsstärke (gering, mittel, hoch) ausgezeichnet. Es ist nicht klar, wie diese Bewertungen zu dem obigen finalen Format übertragen wurden. Diese Original-Version wurde von Clematide und Klenner (2010) auch nochmal über einen mehrstufigen semi-automatischen Prozess erweitert, der hier nur kurz skizziert wird. Es wurde, analog zu anderen Verfahren mit Seed-Words (siehe Kapitel XX), über einen Satz fester Polaritäts-Wörter weitere SBWs in einem Korpus akquiriert, die noch von menschlichen Annotatoren überprüft und bewertet wurden.

# Forschung CD

### German Polarity Clues (GPC)

Das German Polarity Clues (GPC) ist ein deutschsprachiges Sentiment-Lexikon von Waltinger (2010). In der vorliegenden Arbeit wird die Version 0.1 verwendet. Diese wurde in einer Publikation präsentiert und ist ausreichend dokumentiert (Waltinger, 2010). Das Lexikon besteht aus insgesamt 10 141 Wörtern die bezüglich Polarität annotiert sind. Ähnlich zum NRC ist das Lexikon so aufgebaut, dass zu jedem Wort die Polaritätszugehörigkeit über einen dichotomen Wert angegeben wird. Als Polaritätsklassen sind positiv, negativ und neutral enthalten. Eine Auszeichnung mit 1 weist auf eine Zugehörigkeit zur Polarität hin, der Wert 0 besagt, dass das Wort nicht der Polaritätsklasse angehört. Intensitäten sind nicht enthalten. Neben den Polaritätsannotationen wird für manche Wörter die Wahrscheinlichkeit des Auftretens in dem bei der Evaluation des Lexikons verwendeten Korpus angegeben (siehe weiter unten). Nach Analyse der Angaben, hat man sich entschieden die Wahrscheinlichkeiten nicht als Intensitäten zu verwenden, da diese den endgültigen Klassenzugehörigkeiten oft widersprechen, starke Ambiguitäten entstehen und diese Angaben nicht immer vorliegen (unter anderem da manche Wörter nicht im Evaluationskorpus enthalten waren). Das GPC enthält auch flektierte Formen von Wörtern. Zu jedem Wort werden linguistische Informationen wie die Grundform und das POS-Tag angegeben. Es sind alle bedeutenden Wortarten enthalten, auch unübliche Wortarten für Sentiment-Lexika wie z.B. Artiekl. Es ist nicht sicher, ob alle Angaben bezüglich Grundform und POS-Tag vollständig korrekt sind. Daraufin folgen die Angaben zur Polaritätsklasse (positiv, negativ, neutral) und die eben erwähnten Wahrscheinlichkeiten. Folgender Screenshot illustriert den Aufbau:

Das Lexikon enthält 5749 als negativ, 2994 als positiv und 1384 als neutral annotierte Wörter. In der Tat ergibt das 10 127 Wörter, für 14 Wörter liegen keine Annotationen vor, diese sind also für alle drei Polaritätsklassen mit 0 belegt. Ferner sind 110 Wörter für mehrere Klassen annotiert und einige Sonderfälle wie Satzzeichen enthalten. Es sind auch 290 Negations-Phrasen wie z.B. „nicht schlecht“ enthalten. In Kapitel XX werden diese Umstände und der Umgang damit im Detail geschildert. Ein Beispiel für ein positives Wort ist „wunderbar“, für ein negatives „ängstlich“, für ein neutrales „Boden“.

Das GPC wurde mittels eines semi-automatischen Übersetzungsprozesses erstellt (Waltinger, 2010). In einem ersten Schritt wurden einige englischsprachige SA-Lexika bezüglich ihrer Performanz experimentell evaluaiert (Waltinger, 2010b). Die Einträge der besten SA-Lexika dieser Evaluation wurden dann automatisch in die deutsche Sprache übertragen. Es handelt sich bei den Lexika um das SubjectivityClues (Wiebe et al., 2005) und SentiSpin (Takamura et al., 2005). Dabei wurden auch Übersetzungen die mehr als ein Wort ergaben übernommen (maximal drei). Die Polarität wurde aus den Original-Lexika übertragen. Die deutsche Version des SubjectivityClues enthält dabei 9827 Terme, das deutschsprachige SentiSpin 105 561 Terme. Das German Polarity Clues wurde nun durch manuelle Überprüfung aller Terme erstellt. Durch die manuelle Überprüfung wurden beispielsweise Ambiguitäten aufgelöst. Ferner wurden die wichtigsten, bereits oben angesprochenen, Negationsphrasen und einige einschlägige häufige Synonyme enthaltener Terme hinzugefügt. Das finale Lexikon wurde in einer Evaluation noch hinsichtlich seiner Leistung untersucht. Als Test-Korpus wurde ein Korpus bestehend aus Amazon-Produkt-Reviews erstellt, die gemäß der Stern-Bewertung als negativ, neutral und positiv eingestuft wurden. Als Baseline wurden die Leistung der englischsprachigen Original-Lexika auf einem Referenzkorpus von Film-Reviews (Pang et al., 2002) verwendet, die in etwa 83% bezüglich des F-Wertes beträgt. In der Evaluation wurden die deutschsprachigen Versionen des SubjectivityClues und des SentiSpin mit dem eigens erstellten GPC verglichen. Es konnte tatsächlich gezeigt werden, dass das GPC mit einem F-Wert von 87,6% die beste Leistung erbringt. Obschon das Lexikon also im Vergleich dem deutschssprachigen SentiSpin deutlich kleiner ist, erbringt es die bessere Leistung. Ferner wurden in der Evaluation Auftretenswahrscheinlichkeiten der Lexika-Terme für positive, neutrale und negative Reviews erhoben und für das GPC annotiert. Aus den oben genannten Gründen werden diese jedoch nicht weiter verwendet. Das vollständige Evaluationsverfahren wird bei Waltinger (2010) beschrieben.

#Forschung

### Sonstige SA-Lexika

Vereinzelt konnten noch andere deutschsprachige Sentiment-Lexika oder Lexika aus anderen Forschungsgebieten, die jedoch für den Zweck der SA genutzt werden könnten, identifiziert. Diese Lexika wurden aus unterschiedlichen Gründe nicht in der vorliegenden Studie eingesetzt. Zu diesen Gründen gehören die geringe Term-Anzahl, die als geringer eingeschätzte Etablierung in der bisherigen Forschung sowie die als geringer eingeschätzte Eignung. Es wurde auf eine breite Auswahl an Ansätzen und Sentiment-Formaten geachtet. Es konnte kein essentieller Mehrwert durch das Hinzufügen weiterer Lexika für das vorliegende Projekt erkannt werden. Nach Ansicht des Autors sind die wichtigsten und leistungsstärksten vertreten und der lexikonbasierte SA-Ansatz kann erschöpfend für die Dramenanalyse evaluiert werden. Eine mögliche Fehlentscheidung und eine bessere Tauglichkeit anderer Lexika kann jedoch nicht ausgeschlossen werden. Zukünftige Forschung kann sich der Analyse und Evaluation dieser annehmen.

Die wichtigsten Lexika, die nun nicht verwendet wurden, seien hier kurz genannt und beschrieben. Das MLSA (Multi-layered Reference Corpus for German Sentiment Analysis) ist ein Korpus, der für verschiedene Zwecke der deutschsprachigen SA genutzt werden kann. Das Korpus besteht aus mehreren „Schichten“. Auf der obersten Ebene besteht es aus 270 Sätzen, die manuell bezüglich Objektivität, Subjektivität und Polarität annotiert wurden. Auf der zweiten Ebenen wurden Wörter und Phrasen dieser Sätze bezüglich Polarität annotiert. Die dritte Ebene wird als Ausdrucks-Ebene bezeichnet. Hier werden netzartige Bezugsrahmen von persönlichen Zuständen annotiert, die typischerweise aus einer Quelle, also meist einem Sprecher, und einem Bezugsziel sowie anderen interagierenden Faktoren bestehen. Das MLSA kann für verschiedene Anwendungsfälle in der SA genutzt werden. Clematide et al. (2012) empfehlen vor allem die Verwendung als Gold Standard. So kann man beispielsweise mit den eigenen SA-Ansatz mit den annotierten Sätzen der ersten Ebene evaluieren. Emerson und Declerck (2014) evaluieren ihren Ansatz und andere SA-Lexika mit Hilfe der Angaben der zweiten Ebene. Die Annotationssyntax ist dabei komplexer und informationshaltiger als bei den sonstigen Lexika. Bearbeitet man die zweite Ebene und extrahiert die annotierten Wörter zur Erstellung eine SA-Lexika erhält man 777 mit Polaritätsinformationen ausgezeichnete Token (Emerson & Declerck, 2014). Im Vergleich zu anderen hier verwendeten Lexika ist diese Zahl sehr gering.

Bei der Erstellung des GPC wurden durch automatische Übersetzungsverfahren deutschsprachige Versionen der in der englischsprachigen Forschung etablierten (Nachweis) Lexika SubjectivityClues (Wiebe et al., 2005) und SentiSpin (Takamura et al., 2005) erstellt (siehe auch Kapitel XX). Beide Lexika weisen eine überdurchschnittliche Größe auf: das deutschsprachige SubjectivityClues enthält 9827 Terme, SentiSpin 105 561 Terme. Emerson und Declerck (2010) verwenden unter anderem das letztgenannte zur Erstellung eines kombinierten Lexikons. Die Evaluation mit GPC von Waltinger (2010) zeigt jedoch, dass das GPC eine bessere Leistung aufweist. Ferner wurde das GPC durch Kombination und manuelle Ausbesserung der zuvor genannten Lexika erstellt, weswegen eine bessere Performanz von GPC erwartet wird und die beiden anderen Lexika in dieser Studie nicht mitaufgenommen wurden.

In Kapitel XX wird die besondere Bedeutung von komplexeren emotionalen Kategorien jenseits von reinen Polaritäten für literarische Texte erläutert. Aus diesem Grund wurde das NRC als Sentiment-Lexikon für das vorliegende Projekt aufgenommen. Es konnten noch zwei weitere deutschsprachige Lexika identifiziert werden, die derartige Emotionskategorien enthalten. Das Affektive Diktionär Ulm (ADU; Nachweis) stammt aus dem Kontext der Psychologie und Depressionsforschung und umfasst etwa 25 000 flektierte Wörter mit Angaben für 8 Emotionskategorien, von denen einige Unterkategorien, so dass das Lexikon differenzierte Annotationen für 12 Emotionsklassen enthält. Das Lexikon wird grundsätzlich in der therapeutischen Psychologie zur Inhaltsanalyse genutzt, beispielsweise zur Vokabular-Analyse von Patienten (Ortner, 2014, S. 202-203). Aufgrund der Größe des Lexikons und der Betrachtung von unterschiedlichen Emotionen ist ein Einsatz auf literarischen Texten möglicherweise gewinnbringend. Die Nutzung des ADU konnte jedoch nicht im zeitlichen Rahmen des Projekts offiziell beantragt werden.

Das German Emotion Dictionary (Name nach IGGSA) wurde von Klinger, Suliya & Reiter (2016) im Rahmen eines Projekts (siehe Kapitel XX) erstellt. Dazu beziehen sie aus dem SentiWS, dem GPC und dem NRC manuell passende Wörter die zu 7 Emotionskategorien passen und erweitern diese mit Synonymen. Sie erhalten ein Lexikon von 4735 Wörtern, von denen jedes mit einer Kategorie verbunden ist. Das Lexikon ist dabei ungleich bezüglich der Kategorie Verachtung verteilt, die über 2000 Wörter ausmacht. Ferner ist die Genese und Verwendung des Lexikons mangels einer größeren Publikation nur geringfügig dokumentiert. Im vorliegenden Projekt ist die SA auf Polaritätsebene der zentrale Untersuchungsansatz. Auch aus diesem Grund wurde die Verwendung des NRC vorgezogen, da dieses sowohl Polaritäten als auch Emotionskategorien besitzt. Zukünftige Studien, die im Speziellen Emotionen untersuchen wollen, sollten jedoch die beiden letztgenannten Lexika einbeziehen.

Eine systematische Lexika-Auflistung für den deutschsprachigen Bereich ist nicht bekannt. An dieser Stelle seien als weitere Beispiele noch folgende Lexika genannt: Das ANGST (Affective norms for German sentiment terms), das nach einem ähnlichen Konzept wie das Bawl aufgebaut ist, die semi-automatische Übersetzung von SentiStrength (Thelwall et al., 2010) von Momtazi, die Sentiment Phrase List von Rill et al. (2012) sowie SentiMerge, eine Kombination einiger bereits genannter Lexika von Emerson und Declerck (2014). (#Argument gegen SentiMerge).

Insgesamt ist man der Ansicht, dass die letztendlich verwendeten Lexika die wichtigsten für den deutschsprachigen Bereich sind. Es wird ein breites Spektrum an Methoden, Formatierungen und Anwendungsgebieten abgedeckt (Nachweis). Die Frage nach dem Nutzen des Einsatzes von Lexikon-basierten SA-Verfahren in der Dramenanalyse kann über die Auswahl ausreichend und detailliert beantwortet werden.

## Sentiment Analysis in der Literaturwissenschaft

Es werden nun einige Projekte und Studien vorgestellt, die sich mit Sentiment Analysis im Kontext der Literaturwissenschaft befassen, die also Methoden der SA auf literarischen Texten durchführen. Unter literarischen Texten werden im folgenden Texte wie Romane, Märchen, Dramen, Kurzgeschichten und Gedichte verstanden, erzählerische Textformen also, die in einem kreativen Schaffensprozess entstanden sind. Andere literaturwissenschaftliche Abgrenzungen bennenen derartige Textformen auch poetische oder fiktionale Texte in Abgrenzung zu Sachtexten (Frederking, 2016). Die klare definitorische Abgrenzung zu anderen Textformen, die in großem Umfang in der SA untersucht wurden, wie z.B. Blog-Posts, ist natürlich nicht immer deutlich. Eine tiefergehende literaturwissenschaftliche Diskussion dieser Abgrenzung ist nicht Teil dieser Arbeit; durch die Definition der obigen Gattungen als literarische Texte ist die Abgrenzung gegenüber Produkt-Reviews, Social Media, Blog-Posts oder News-Artikeln für die SA ausreichend. Für die vorliegende Arbeit besonders wichtig, sind dabei Projekte, die sich mit der SA in Dramen befassen.

Kakkonen und Kakkonen (2011) stellen fest, dass im Vergleich zur Forschung auf Anwendungsgebieten wie Produkt-Reviews oder Twitter (siehe Kapitel XX) Sentiment Analysis bislang kaum auf literarischen Texten angewandt und untersucht wurde. Marchetti, Sprugnoli und Tonelli (2014) erweitern diese Aussage, indem sie einen Mangel an Forschung bezüglich SA in den Geisteswissenschaften generell feststellen. Betrachtet man Ergebnisse von Meta-Studien zur SA wird diese Aussage vollständig bestätigt (Nachweis). In Kapitel XX kann man die zentralen Themengebiete und Untersuchungsgegenstände der SA-Forschung nachlesen. Dennoch konnten einige Projekte und Studien identifiziert werden, die erste SA-Möglichkeiten explorieren. In der Tat sind Emotionen und Gefühle essentielle Bestandteile zur Analyse und zum Verständnis von literarischen Texten (Mohammad, 2011; Nalisnick & Baird, 2013). Das grundlegende Ziel des Forschungsbereichs ist es die literaturwissenschaftliche Arbeit und die hermeneutische Textanalyse mittels SA zu unterstützen und zu erweitern. Neben der Analyse und Exploration von konkreten SA-Methoden auf literarischen Texten werden vor allen Dingen Möglichkeiten und Probleme der Erstellung annotierter Korpora, die für fortgeschrittene SA-Verfahren notwendig sind, in Studien untersucht.

Zunächst wird die Arbeit von Alm und Sproat (2005) vorgestellt. Sie befassen sich mit emotionalen Trends und Verläufen in Märchen. Die Arbeit ist dabei Teil eines größeren Projekts mit dem Ziel die Prädiktion von Emotionen in Texten zur Anpassung der emotionalen Tonlage in einem auf Kinder ausgerichtetem Sprachsynthese-System zu nutzen. Die erste Studie beschäftigt sich mit der Annotation von Märchen bezüglich Emotionen und es werden noch keine konkreten SA-Methoden angewandt. Die Annotation literarischer Texte von menschlichen Beurteilern wird jedoch auch in der vorliegenden Arbeit durchgeführt, weswegen das Vorgehen und die Ergebnisse trotzdem bedeutend für Teile des hier vorgestellten Projektes sind. Als Korpus dienen 22 Märchen der Brüder Grimm. Insgesamt haben für 2 Sets von Märchen je zwei Annotatoren unabhängig voneinander jeden Satz bezüglich 8 Emotionskategorien ausgezeichnet und also angegeben ob und welche Emotion vorliegt (wütend, angeekelt, furchtsam, glücklich, traurig, positiv überrascht, negativ überrascht). Die Annotatoren haben einen Literaturkurs zu den Märchen von Grimm besucht und wurden in die Aufgabe eingeleitet. Die Übereinstimmungsergebnisse bei der unabhängigen Auszeichnung ergeben jedoch eher geringe Übereinstimmungswerte, sowohl über Cohens Kappa (Für Gruppe 1 und Set 1: 0,51; für Gruppe 2 und Set 2: 0,24) als auch über die prozentualen Übereinstimmungen zwischen den Annotatoren (0,64; 0,45). Alm und Sproat (2005) weisen selber darauf hin, dass diese Ergebnisse deutlich schlechter sind als emotionale Annotationen in anderen Bereichen. Diese Einschätzung stimmt mit den in Kapiel XX beschriebenen Ergebnissen überein. Die Werte verbessern sich, wenn man statt den Einzelgruppen, Oberkategorien betrachtet, z.B. neutral oder nicht-neutral, sowie neutral oder positiv oder negativ. Die prozentuale Übereinstimmung für das erste Set liegt dann bei 0,76 und 0,73, was näher an anderen Ergebnissen in der SA liegt. In der Detailanalyse kann man feststellen, dass Annotatoren sich stark unterschiedlich verhalten, manche zeichnen beispielsweise sehr viele Sätze neutral aus, manche gar keine. Um die Probleme noch weiter hervorzuheben lassen sie 3 Annotatoren noch mal 4 Märchen erneut für die Emotionskategorien auszeichnen. Hierbei zeigt sich zwar eine bessere Übereinstimmung der Annotatoren mit sich selbst als bei den vorherigen Messungen zwischen den Annotatoren (0,60-0,68), die dennoch noch entfernt von einer fast vollständigen Übereinstimmung sind. Es wird also deutlich, dass auch außer-textliche Faktoren wie möglicherweise die Gemütslage die Auszeichnung beeinflussen. Auch hier werden die Übereinstimmung deutlich besser, wenn man statt der Übereinstimmung der Emotionsangaben Oberkategorien analysiert. Die Ergebnisse der Annotation zeigen die besondere Schwierigkeit der Annotation von literarischen Texten im Vergleich zu Textformen wie Produkt-Reviews und Tweets, und wie diese leicht unterschiedlich interpretiert werden können. Dennoch wurden die Daten für eine weitere Analyse verwendet. Die Mehrheitsangaben der Annotatoren wurde dabei als endgültiger Wert entnommen. Bei Gleichheiten hat ein Experte entschieden. Auf diese Weise konnten erste Analysen bezüglich Emotionsverteilungen und -verläufen in Märchen durchgeführt werden. Etwa 60% der Sätze sind neutral, die häufigste annotierte Emotion ist mit 12% Zorn. Auffällig ist die Verteilung, wenn man zwischen positiven und negativen Emotionen unterscheidet, denn lediglich 10% sind positiv und 30% negativ. Dabei ist jedoch zu beachten, dass es lediglich 2 Emotionskategorien gibt, die als positiv eingestuft wurden. Über statistische Satzanalyse konnten literaturwissenschaftliche Fragestellungen beantwortet werden. Die Märchen beginnen häufiger neutral und enden meisten glücklich. Neutrale Sätze sind für die meisten Emotionssätze der Kontext. Die Emotionen Wut und Traurigkeit werden meist von Sätzen der gleichen Kategorie umgeben, während dies nicht für die anderen Kategorien gilt. Das zeigt, dass manche Emotionen wie Überraschung und Furcht plötzlich auftreten im Gegensatz zu den anderen genannten Emotionen, die meist Teil einer längeren Erzählung sind. Die Märchen werden in fünf Akte zerteilt und die emotionale Verteilung betrachtet und versucht im Kontext literaturwissenschaftlicher Fragestellungen zu interpretieren. Der Anstieg negativer Sätze im letzten Akt lässt sich so möglicherweise über die typische Enthüllung des Schurken erklären.

Die Arbeit von Alm und Sproat (2005) ist Teil eines größeren Projekts und im selben Jahr veröffentlichen Alm, Roth und Sproat (2005) einen weiteren Beitrag, in dem sie nun, analog zu Ansätzen in der SA, maschinelles Lernen zur Prädiktion von Emotionen auf Satzebene für die Textsorte Märchen anwenden. Es handelt sich dabei um den gleichen Korpus und das gleiche Annotationsschema (7 Emotionskategorien) wie oben beschrieben. Sie erläutern auch ihren Fortschritt bei der manuellen Annotation des Gesamt-Korpus aus 185 Kindermärchen. Die Annotatoren-Übereinstimmung ist weiterhin sehr gering. Als Hauptproblem sehen sie die Entscheidung zwischen dem generellen Vorhandensein einer Emotion und deren Abwesenheit, also den Umgang mit Neutralität. Mit dem selben Test-Korpus, wie in der vorigen Studie bereits beschrieben, aus 22 Märchen und 1580 Sätzen nutzen sie einen linearen Klassifikator im Vektor-Raum um den Einsatz von maschinellen Lernen für die Prädiktion von Emotionen in Märchen zu ergründen. Dabei testen sie Kombinationen von verschiedenen Feature-Sets. Als Features dienen dabei linguistische Aspekte der Sätze wie z.B. die Anzahl der Wörter oder die Verteilung der Wortarten aber auch die Zahl von positiven, negativen und Emotionswörtern. Letztere werden durch verschiedene Lexikonlisten kalkuliert, u.a. die Emotionswörter von WordNet (Fellbaum, 1998). Ferner wurde der Einfluss von „sequencing“, also der emotionalen Annotation des Vorgänger- und Nachfolgersatzes untersucht. Die besten Prädiktionsergebnisse konnten durch die Integration aller Merkmale und „sequencing“ erreicht werden. Je nach Evaluationsverfahren weist eine 10-fache Kreuzvalidierung Erkennungsraten von bis zu 69% für die Prädiktion der binären Gruppe: Emotion vorhanden oder neutral. Damit liegt die Erkennungsrate über der Mehrheits-Baseline von 60% (also die Leistung bei der grundsätzlichen Zuweisung der häufigsten Klasse neutral). Bei der Ausweitung der Vorhersage auf die Gruppe Neutralität, Positivität und Negativität werden die Ergebnisse vor allem für Vorhersage der Polaritäten sehr schlecht. Diese haben F-Werte von lediglich 0,32 (negativ) und 0,13 (positiv). Dies liegt aber auch am kleinen Trainingsset, da 60% der Daten neutral annotiert sind und lediglich 10% überhaupt positiv. Insgesamt konnten Alm et al. (2005) erstmal jedoch, unter dem Hintergrund der Herausforderungen des Untersuchungsgegenstands in der Annotation und Interpretation, den gewinnbringenden Einsatz von überwachten Lernalgorithmen zur SA auf literarischen Texten belegen. Der erfolgreiche Einsatz von Wortlisten aus Sentiment-Lexika als Features für den ML-Algorithmus zeigt auch den Nutzen dieser Methodik auf. Im Rahmen des Sprachsynthese-Projekts explorieren Alm und Sproat (2005b) auch die Interpretation und Wahrnehmung von Emotionen auf phonetischer Ebene. Die Studie hat aber in seinen Ansätzen und Zielen nur noch entfernt mit SA auf Textebene zu tun und hat damit keine größere Bedeutung für die vorliegende Arbeit.

Volkova et al. (2010) greifen die Erfahrungen bezüglich der Annotation von Alm und Sproat (2005) auf und explorieren auch das Annotationsverhalten bezüglich Sentiments. Die Studie bringt für die vorliegende Arbeit aufgrund der deutschsprachigen Ebene des Korpus und der Annotation einen Mehrwert. Zehn deutsche Muttersprachler nahmen an der Studie teil. In einigen Nebenstudien konnten Informationen über die Intensität, Polarität und das Clustering der Emotionskategorien erhoben werden. Speziell für das Deutsche konnte somit zum Beispiel festgestellt werden, dass im Gegensatz zum Englischen die Emotion Überraschung klar positiv und nicht ambivalent ist (Alm & Sproat, 2005). Zentraler Bestandteil ihrer Studie ist jedoch die Annotation von 8 Märchen der Brüder Grimm. Die Annotationsanweisung war dabei speziell und entfernt sich von Anweisungen bei ähnlichen Studien: Teilnehmer sollten diejenigen Teile des Textes markieren bei der sich eine Emotion in Stimmlage oder Gesichtsausdruck äußern würde, wenn man den Text laut vorläse. Volkova et al. (2010) gehen von einer sehr breiten Emotionsspanne aus und wählen je 7 positive (z.B. Mitgefühl) und 7 negative Emotionskategorien (z.B. Hass) sowie die Klasse neutral für die Annotation aus. insgesamt konnten so insgesamt 150 Texte von Annotatorenpaaren ausgezeichnet werden (zwei Texte von zehn, sechs Texte von fünf Annotatoren). Die Resultate zeigen, dass Teilnehmer meist für Phrasen von 4-7 Token eine emotionale Äußerung identifizieren. Analog zu Alm & Sproat (2005) zeigt sich, dass die Mehrzahl des Textes als neutral bzw. nicht annotiert wurde (etwa 65-75%). Entgegen zu Alm und Sproat (2005) stellen sie insgesamt ein Übergewicht von positiven Emotionen fest. Dies hängt aber auch stark vom Text abhängt, vereinzelte Märchen werden eher sehr negativ annotiert. Auch bezüglich des Emotionsverlaufs im Drama können Volkova et al. (2010) zeigen, dass die Menge und Stärke an Emotionswörtern (Wörtern und Phrasen, die als emotionstragend annotiert wurden) insgesamt am Anfang stark ansteigt, dann im Verlauf abfällt um am Ende wieder anzusteigen. Bezüglich der Annotationsübereinstimmung stellen sie eine insgesamt moderate Übereinstimmung (0.53) für manche Paare eine fast perfekte fest unter Betrachtung zusammengehöriger Emotionen in Emotionsclustern. Werden alle Emotionen einzeln betrachtet ist die Übereinstimmung im Durchschnitt knapp moderat aber deutlich schwächer (0,34). In einer der zuvor erwähnten Vorstudien annotierten Teilnehmer Wörter aus dem Märchen-Vokabular kontextfrei bezüglich Polarität. Diesbezüglich kann eine schwächere Übereinstimmung festgestellt werden von 0,46, was aber von Volkova et al. (2010) auf das häufige Auftreten der Neutralität zurückgeführt wird. Insgesamt können Vokova et al. (2010) über ihre Methodik ähnliche Übereinstimmungen erlangen wie Alm und Sproat (2005), für manche Annotatorenpaare auch sehr hohe Übereinstimmungen und dies obschon die Annotation deutlich differenzierter durchgeführt wurde und Differenzierungen eher zu mangelnder Übereinstimmung bei Alm und Sproat (2005) führten. Ein Auswertungsschritt relativiert jedoch die Ergebnisse und ist kritisch zu betrachten. So entfernen sie Annotatorenpaare, die sehr geringe Übereinstimmungsstatistiken haben und die wenig Text als emotional markierten. Alm und Sproat (2005) können derartiges Verhalten auch identifizieren und sehen das als Zeichen des breiten individuellen Interpretationsspielraums der gerade bei literarischen Texten besteht. Insgesamt zeigen auch Volkova et al. (2010) die Probleme der Annotation literarische Texte auf.

Kakkonen und Kakkonen (2011) entfernen sich vom reinen Annotationsproblem und entwickeln ein SA-Tool SentiProfiler mit dem man Sentiment-Profile von Texten vergleichen kann. Dazu definieren sie zunächst das Konzept von Sentiment-Profilen als hierarchische Repräsentationen des affektiven Inhaltes von Dokumenten. Zur Bestimmung dieser Sentiment-Profile wird unter anderem ein Lexikon-basierter SA-Ansatz über das WordNet-Affect (Strapparava & Valitutti, 2004) umgesetzt. Das WordNet-Effect ist eine linguistische Ressource, die eine hierarchische Struktur für Emotionen definiert und Wörter für jede der Emotionskategorien angibt. Auf diese Weise kann automatisch eine Emotions-Ontologie erstellt werden, die auch zur Generierung der Sentiment-Profile genutzt wird. Der „negative-emotion“-Zweig der Ontologie enthält beispielsweise 147 Unterklassen mit einer maximalen Knotentiefe von 5. Dieser Zweig besteht aus 832 Wörtern, die den einzelnen Kategorien zugeordnet sind. Über einen Wortabgleich der Ontologie-Wörter mit dem eingegebenen Text kann bestimmt werden, welche Klassen in welchem Ausmaß den Text bestimmen. Das Sentiment-Profil ist die visuelle Repräsentation des Vorkommens der Klassen als hierarchische Graph-Struktur mit den Klassen als Knoten. Zu jedem Knoten wird die Zahl der im Text zu der Klasse gehörenden auftretender Wörter angegeben. Ferner wird ein normalisierter Wert definiert und angegeben: Die Zahl aller Wörter einer Klasse geteilt durch alle Wörter. Auf Basis der Graph-Struktur können für Oberknoten aggregierte Werte angegeben werden. Es wird auch eine Vergleichs-Visualisierung für zwei Eingabetexte implementiert, bei der Klassen die in einem Text vorkommen, aber nicht im anderen durch zusätzliche Knoten angezeigt werden und Differenzen bei gemeinsamen Klassen angezeigt werden. Über Farben werden diese Unterschiede noch verdeutlicht.

Kakkonen und Kakkonen (2011) nutzen diese SentimentProfile um die Unterscheidung der „Terror“- und „Horror“-Genre in Schauerromanen (gotischen Romanen) zu untersuchen. Es handelt sich dabei um eine Forschungsfrage aus der Literaturwissenschaft (Botting, 1996). Kakkonen und Kakkonen nutzen dabei nur den oben beschriebenen „negative-emotion“-Zweig und vergleichen zwei typische Romane des jeweiligen Genres. In der Tat können sie beispielsweise zeigen, dass, im Vergleich, der Horror-Roman größere Werte bezüglich den Klassen Ekel und Übelkeit enthält, während der Terror-Roman im Vergleich mit weniger intensiven Klassen zusammenhängt wie Ungeduld und Depression. Ähnliche Ergebnisse können sie auch für äquivalente Romanvergleiche identifizieren. Kakkonen und Kakkonen (2011) sehen damit literaturwissenschaftliche Interpretationsansätze bestätigt. Insgesamt exploriert das Projekt von Kakkonen und Kakkonen einen fortgeschrittenen SA-Ansatz über die Definition sehr komplexer hierarchischer Sentiment-Profile, die weit über die herkömmliche Polaritätsbestimmung in der SA hinausgeht. Obschon sie die gewinnbringende Nutzung für einen Anwendungsfall aufzeigen werden die Graph-Strukturen sehr groß und es besteht die Gefahr von Unübersichtlichkeit obwohl man sich nur auf eine Oberkategorie beschränkt hat. Dennoch kann man so einen sehr differenzierten Blick auf die emotionale Verteilung eines literarischen Textes gewinnen.

Ein weiteres Projekt, dass in vielen Aspekten nah an der vorliegenden Arbeit liegt, stammt von Mohammad (2011). Es handelt sich dabei auch um einen der Ersteller des NRC Emotion Association Lexicon (Mohammad & Turney, 2010). Mohammad nutzt dieses Lexikon zur SA von Märchen, Dramen, Romanen und einem Google-Books-Korpus um Möglichkeiten zur Analyse und Visulisierung der Sentiment-Informationen zu explorieren. Das verwendete Lexikon (NRC) wurde bereits in Kapitel XX beschrieben. Er akquiriert verschiedene adäquate Korpora für die jeweiligen Gattungen. Über eine simple Ad-Hoc-Verwendung des Lexikons kalkulieren sie somit für die einzelnen Texteinheiten emotionale Verteilungen indem er die Zahl der Wörter einer Kategorie an der Gesamtzahl von Wörtern teilt, die eine emotionale Kategorie besitzen. Dadurch erhält er die prozentuale Verteilung der Emotionen in einer Texteinheit gemessen an allen Emotions-Wörtern. Er visualisiert diese Verteilung mit Kreisdiagrammen für zwei Shakespeare-Dramen. Über Subtraktion der Verteilungswerte zweier Shakespeare-Dramen kann er Unterschiede in der Verteilung über ein Balkendiagramm aufzeigen. So kann er zeigen, dass das Drama Hamlet einen höheren Anteil an Furch, Traurigkeit, Ekel und Wut hat jedoch weniger Freude, Vertrauen und Erwartung aufweist als das Drama „Wie es euch gefällt“. Er definiert die relative Salienz eines Wortes als Metrik um die zentralen Wörter in zwei Zieltexten. Es handelt sich um die Differenz der Worthäufigkeit normalisiert an der Zahl aller Wörter pro Text. Somit kann er in Word-Clouds die Wörter visualisieren die insbesondere zu Unterschieden zwischen zwei Texten führen. Über Liniendiagramme visualisiert er unter anderem den Verlauf von Emotionsverteilungen entlang von Dramensegmenten. Er definiert den Begriff der Emotionsdichte als erwartete Zahl von Emotionswörtern pro X Wörter einer Texteinheit. Er wählt für X die Zahl 10 000. Es handelt sich also um eine statistische Normalisierung entlang eines festen Wertes, man multipliziert die Zahl der Emotionswörter mit dem Faktor, mit dem man die Gesamtzahl der Wörter des Textes multiplizieren muss, um X zu erhalten. Er visualisiert z.B. die Emotionsdichte für negative Wörter für Märchen der Brüder Grimm in einem Liniendiagramm. Ferner analyysiert er die Emotionsverteilung eines großen Google-Book-Korpus entlang der Jahresachse und für verschiedene Nationen. So erkennt er zum Beispiel einen Anstig von Furch-Wörtern in Büchern während des ersten Weltkriegs in Deutschland. Auch hierfür nutzt er Liniendiagramme. In einer abschließenden Fragestellung versucht er die wichtigen literaturwissenschaftlichen Unterschiede zwischen Romanen und Märchen anhand seiner Metriken herauszuarbeiten. Er kann anhand Emotionsdichten zeigen, dass der Märchen-Korpus signifikant mehr Erwartungs-, Freude-, Ekel- und Überraschungsdichten haben als Romane. Sie haben jedoch weniger Vertrauensdichten. Des Weiteren haben Märchen signifikant weniger negative Terme und mehr positive. Ferner haben mehr Märchen sehr hohe und sehr geringe Emotionsdichten als Romane. Insgesamt kann Mohammad (2011) zeigen wie man mit sehr einfachen Mitteln der SA Impulse für die literaturwissenschaftliche Arbeit liefern kann. Er verdeutlicht aber auch die Herausforderungen, die die SA speziell in diesem Themenbereich hat. Es müssen spezielle Metriken und Visualisierungen angepasst an das Themengebiet entwickelt werden. Dieser Ansatz wird in der vorliegenden Arbeit aufgegriffen.

Elsner (2012) beschäftig sich mit der Repräsentation des Handlungsverlaufs (Plot-Struktur) in Romanen und nutzt dabei neben anderen Ansätzen auch Methoden aus der SA. Bezüglich der SA-Komponente seiner Arbeit, ist sein Ziel die Erfassung der Emotionen einer Figur und des emotionalen Zusammenhangs von paarweisen Figurenbeziehungen. Dazu zerlegt er die Romantexte seines Korpus aus 11 Liebesromanen des 19. Jahrhunderts in Paragraphen. Je nachdem ob exakt eine oder zwei Figuren in einem Paragraph erscheinen wird der Text als der Figur bzw. der Figurenbeziehung zugehörig betrachtet und eine simple SA nach Lexikon-basierten Verfahren und Term-Zähl-Methodik durchgeführt. Als Lexikon verwendet er einen von Wilson et al. (2005) erstelltes Korpus aus subjektiven Ausdrücken, die als emotionsbeladen annotiert sind. Er betrachtet also die Emotionalität der mit Figuren assoziierten Sprache über den Romanverlauf. Über Normalisierungen und Interpolationen kann er mit Hilfe dieser Heuristik grafisch auch den emotionalen Verlauf einer Figur im Roman visualisieren. Diese SA-Komponente ist dabei Teil eines größeren Programms, um Romane bezüglich ihrer Plot-Struktur mit Hilfe der Romanfiguren miteinander zu vergleichen. Er belegt die Funktionalität und den Nutzen des Programms indem er einen ML-Algorithmus (der unter anderem die Emotionsinformationen der Figuren nutzt) implementiert mit dem er effektiv richtige Romane von künstlich erstellten unterscheiden kann. Insgesamt zeigt das Projekt von Elsner wie eine verhältnismäßig simple SA-Herangehensweise erfolgreich in einem größeren Kontext auf literarischen Texten eingesetzt werden kann.

Nalisnick und Baird (2013) nutzen SA nun wieder zur Analyse von Dramen, diesmal jedoch mit einem Fokus auf Figurenbeziehungen. Nalisnick und Baird nennen dies das Character-to-Character Sentiment. Als Korpus dient ein XML-Korpus von Shakespaere-Dramen. Auch hier wird ein simpler lexikon-basierter Ansatz verfolgt. Als Lexikon wird das AFFINN Sentiment Lexicon von Nielsen (2011) genutzt. Es gibt Valenzwerte auf einer absoluten Skala von -5 (sehr negativ) bis 5 (sehr positiv) an. Es besteht aus 2477 Wörtern und Phrasen. Als Heuristik für ein Character-to-Character Sentiment wird angenommen, dass sich jede Replik eines Sprechers auf den zuvor sprechenden Sprecher richtet. Dies ist natürlich häufig nicht korrekt, da sich Sprecher nicht immer in einem klaren aufeinander gerichteten Dialog befinden, oft ist das Sentiment-Target auf das sich der Sprechakt richtet nicht in der Szene oder ein anderer Sprecher. Des Weiteren können Repliken auch sich gar nicht auf irgendwelche Figuren beziehen, sondern andere Inhalte haben. Dennoch können Nalisnick und Baird (2013) mit dieser Heuristik erwartungskonforme Ergebnisse produzieren, weswegen sie auch in der vorliegenden Arbeit aufgegriffen wird. Bezüglich des SA-Vorgehens wird auch hier das Lexikon Ad-Hoc ohne größer Verarbeitungsschritt auf lexikalischer oder textueller Ebene verwendet. Es werden die absoluten Valenzwerte durch einfach Summierung der Polaritätsangaben pro Texteinheit berechnet. Bereits durch diesen simplen Ansatz können sie zeigen für das Drama Hamlet zeigen, dass die an Claudius gerichteten Repliken insgesamt den negativsten Wert bilden, was die Handlung des Dramas korrekt widerspiegelt. Andererseits erkennen sie, dass Gertrude gemäß der Heuristik von Hamlet eine sehr positive Bewertung erfährt, was wiederum nicht konform mit der gängigen Interpretation ist (Nalisnick & Baird, 2013). Zur besseren Analyse visualisieren sie den Valenz-Verlauf über Visualisierungen. Dazu tragen sie auf der x-Achse die Repliken-Zeilen ab und auf der y-Achse die absoluten fortschreitend summierten Werte des Character-to-Character Sentiments als Verlaufslinie. Auf diese Weise können sie zusammengehörige Charakterbeziehungen direkt im dynamischen Verlauf des Stücks vergleichen. So können sie für Gertrude und Hamlet einen drastischen Polaritätswechsel für eine spezielle Szene identifizieren, und können dies erfolgreich mit der Handlung des Dramas interpretieren. Sie können mittels dieser Berechnungen und Visualisierungen für weitere bekannte und Figurenbeziehungen aus Shakespeares Dramen erwartbare Ergebnisse produzieren (z.B. Othello und Desdemona; Julia und Romeo). Insgesamt können Nalisnick und Baird (2013) schon mit einem sehr einfachen SA-Ansatz Figurenbeziehungen sowohl statisch als auch dynamisch analysieren.

In einem sehr aktuellen Beitrag versuchen Jannidis et al. (2016) von der Uni Würzburg „Happy Endings“ in Romanen mit Hilfe von SA vorherzusagen. Es handelt sich also nicht direkt um SA zum Ziel der Polaritätsbestimmung, SA ist hier nur eine Methodik. Da der Untersuchungsgegenstand literarische Texte sind und SA eingesetzt wird, wird das Projekt hier trotzdem kurz beschrieben. Als Testkorpus wird ein Datensatz aus 212 deutschsprachigen Romanen aus dem 19. Jahrhundert verwendet. Für jeden Roman wurde manuell annotiert ob ein Happy End vorliegt oder nicht, was für die Hälfte des Korpus der Fall ist. Zur Vorhersage verwenden sie einen Machine-Learning-Ansatz mittels SVM. Die Trainings-Features generieren sie nun über einen lexikon-basierten SA-Ansatz. Als Lexikon verwenden sie die deutsche Übersetzung des schon mehrfach angesprochenen NRC (siehe Kapitel XX und XX). Jeder Roman wurde in Segmente zerlegt und der Text lemmatisiert. Für jedes Segment wurden über Wortabgleich mit dem Lexikon Durchschnitts-Werte für die Positivität, Negativität und die 8 Emotionskategorien berechnet. Die Gesamt-Polarität wurde durch Subtraktion des Negativität-Werts vom Positivität-Werts bestimmt. Diese 11 Sentiment-Werte wurden als Feature-Set für die SVM verwendet. Nutzt man nur die Werte des letzten Segments erreicht somit eine Erkennungsgenauigkeit für Happy Ends von 67%. Durch eine Anpassung der ausgewählten Segmente und Features kann eine Erkennungsgenauigkeit von 73% der SVM erreicht werden. Durch weitere Analysen können sie ferner feststellen, dass ältere Romane besser erkannt werden als neue. Sie erklären sich diesen Befund mit der stärker schematischen Struktur der vor-realistischen Romane.

Marchetti et al. (2014) explorieren Möglichkeiten der SA auf historischen Texten des italienischen Politikers Alcide De Gasperi (1881 – 1954). Das Korpus besteht aus 3000 Dokumenten und 3 000 000 Wörtern. Obschon es sich nicht um literarische Texte im engeren Sinne handelt ist das Projekt für die vorliegende Studie aufgrund ähnlicher Grundprobleme, die durch die spezielle Textauswahl entstehen, relevant: der Mangel an etablierten nicht-englischen SA-Lexika und das Fehlen eines annotierten Korpus. In einem ersten Versuch sollte die Leistung des Einsatzes von zwei Sentiment-Lexika auf der historischen Domäne überprüft werden. Als SA-Lexika wurden SentiWordNet (Baccianella & Sebastiani, 2010) und WordNet-Affect (Strapparava & Valitutti, 2004) gewählt. Die in den Lexika angegebenen Synsets (also Konzepte mit dazugehörigen Termen, siehe Kapitel XX) basieren au WordNet. Über die Ressource MultiWordNet sind italienische Synsets mit englischen verbunden. Durch diese Übertragung können die genannten Lexika für das Italienische benutzt werden und die italienischen Synsets mit einer positiven, negativen oder neutralen Polarität ausgezeichnet werden. Für die spätere SA wurden dabei die Synsets von beiden Lexika verwendet. Das Korpus wurde lemmatisiert und mit den Polaritäts-Scores der Lexika abgeglichen. Die Scores wurden nochmal manuell überprüft. Die generellen Polaritäten der einzelnen Texte wurden durch ein einfaches Term-Zähl-Verfahren kalkuliert. Über eine simple Visualisierung wurde Feedback von Historikern erhoben. Tatsächlich gaben die Historiker jedoch an, dass sie mehr an der Polarität eines Themas und der zeitlichen Entwicklung dieses Themas interessiert sind. In einem weiteren Experiment wurden themenspezifische Sätze (525) aus dem Korpus akquiriert und bezüglich Polarität annotiert. Die Polarität bezog sich dabei auf das im Satz vorkommende Thema. Dazu wurde jeweils der Vorgänger- und Nachfolger-Satz zur Kontextualisierung mit angegeben. Die Experten erstellten dann einen Gold-Standard mit 60 Sätzen, bezüglich derer sie übereinstimmten. Des Weiteren wurden noch von einer Crowd-Sourcing-Plattform Bewertungen von Nicht-Experten bezogen. Die kontextuelle Polarität jedes Themas in jedem Satz wurde dabei mindestens fünf Mal, gemäß folgender Skala bewertet: positiv, negativ, neutral, unbekannt. Die Annotationen des Crowdsourcing-Ansatzes basieren auf der Mehrheitsentscheidung, also was die meisten Bewerter wählten wurde einem Satz zugeschrieben. Die Bewertungsleistung der Crowd gegenüber dem Gold-Standard wurde als Benchmark für die Lexikon-basierte SA betrachtet. Die Leistung der Crowd war eher durchschnittlich (68%) und verdeutlich die Komplexität des Themas. Die Leistung der SA jedoch war mit 43% unter dieser Mindestbenchmark, und besondere Problem lage bei der Vorhersage negativer und neutraler Sätze vor, d.h. die SA hat überdurchschnittlich stark Sätze positiv bewertet. Ferner wurde die Übereinstimmung von Crowd und Experten analysiert. Auch diese ist sehr gering mit einem Kappa-Wert von 0.39. Insgesamt verdeutlicht die Studie von Marchetti et al. (2014) wieder die Komplexität der Aufgabe der SA auf speziellen Textformen wie hier historischen Texten.

## Quantitative Dramenanalyse

Die SA als sowohl textanalytische und mathematisch-quantitative Methodik wird in ihrem Einsatz auf Dramentexten in der vorliegenden Arbeit als Teilgebiet der quantitativen Dramenanalyse betrachtet.

#TODO

# Forschungsfragen und -Agenda

Im nachfolgenden Kapitel werden die Motivation und die Ziele des Projekts zusammengefasst und beschrieben. Nachdem die grundsätzliche Motivation erläutert wird, wird eine grundsätzliche, übergeordnete Forschungsfrage als Leitfrage formuliert. Daraufhin wird eine Forschungsagende mit konkreten Arbeitsschritten beschrieben, bei der Besonderheiten und bisherige Schwächen des Forschungsgebiets aufgegriffen um mit der vorliegenden Arbeit einen deutlichen Mehrwert für die bestehende SA-Forschung auf literarischen Texten, insbesondere in der Dramenanalyse, zu liefern.

Die grundsätzliche Motivation für die vorliegende Arbeit ist begründet in den allgemeinen Zielen der quantitativen Dramenanalyse und Literaturwissenschaft: die Unterstützung und Erweiterung der literaturwissenschaftlichen Interpretation und Analyse durch computergestützte Methoden. Die Beschäftigung mit Emotionen, Meinungen und Gefühlen in literarischen Texten ist relevant für die Interpretation dieser und die literaturwissenschaftliche Arbeit mit den jeweiligen Texten. Mohammad (2011) weist darauf hin, dass literarische Text wie Romane und Märchen schon immer Kanäle waren um Emotionen implizit und explizit zu transportieren. Nalisnick und Baird (2013) sehen es als zentral für die Leseerfahrung an, Empathie für die emotionalen Hochs und Tiefs von Figuren in literarischen Texten, fühlen. Tatsächlich sind Emotionen und Gefühle essentieller Bestandteil in der literaturwissenschaftlichen Interpretation und können einen hohen Wert für diese einnehmen (Winko, 2003; Mellmann, 2007). Auch bezogen auf den hier vorliegenden Lessing-Korpus findet man Interpretationen und Analysen die sich mit Emotionen und Gefühlen in den Dramen auseinandersetzen (z.B. Alt, 1994, S. 191-210; Fick, 2000, S. 334-335). Es ist also insgesamt naheliegen, dass jede Methodik, die die Analyse von Emotionen und Gefühlen in literarischen Texten vereinfacht oder erweitert, aufgrund der hohen Bedeutung von Emotionen in diesen Texten, einen Mehrwert für die Literaturwissenschaft liefern kann. Die zentrale Methodik in der Informatik zur Analyse von Emotionen und Gefühlen in geschriebenem Text ist die SA. Mit der vorliegenden Arbeit soll also untersucht und diskutiert werden ob die computergestützte Methodik der SA einen Mehrwert für die literaturwissenschaftliche Arbeit leisten kann. Es soll exploriert werden ob bestehende Interpretationen durch SA bestätig oder erweitert werden können und ob die SA neue Einblicke für die literaturwissenschaftliche Arbeit liefern kann. Erste Arbeiten können einen möglichen Nutzen auch bereits punktuell aufzeigen (Mohammad, 2011; Nalisnick & Baird, 2013; Klinger et al., 2016). Grundsätzlich sind Arbeiten, die sich mit der SA in literarischen Texten befassen im Vergleich mit anderen Gebieten noch selten. Ein Projekt, dass den Einsatz von SA speziell in deutschsprachigen Dramen exploriert ist bislang unbekannt. Die hier entwickelten Tools sowie die anderen Bestandteile der Arbeit basieren auf der Motivation und dem Ziel die Literaturwissenschaft und im speziellen in der vorliegenden Arbeit die literaturwissenschaftliche Dramenanalyse bei ihren Aufgaben zu unterstützen und neue Impulse zu liefern.

Auf Basis der eben beschriebenen Ziele und Motivation wird die zentrale übergeordnete Forschungsfrage formuliert:

*Lässt sich eine* ***optimierte Sentiment Analysis*** *zur* ***unterstützenden Beantwortung und Diskussion literaturwissenschaftlicher Interpretationen*** *in der Dramenanalyse einsetzen?*

Die vorliegende Arbeit kann zum gegenwärtigen Zeitpunkt der Forschung die Frage nicht vollständig klären, vielmehr ist die hier formulierte Forschungsfrage als grundsätzliche Orientierung und Leitfrage zu verstehen. Es handelt sich um keine systematisch falsifizierbare Hypothese. Die finalen durchgeführten Arbeitsschritte leisten vielmehr einen Beitrag zur Diskussion und Analyse der Forschungsfrage. Die einzelnen Bestandteile der Forschungsfrage werden nun im Detail betrachtet und eine Forschungsagenda zergliedert, bei der Bezug auf die Probleme der bisherigen Forschung genommen wird, um den Mehrwert des Projekts zu verdeutlichen.

Unter einer *optimierten SA* wird ein SA-Verfahren verstanden, dass die spezifischen Probleme des Anwendungsgebiets aufgreift, versucht Verbesserungen herbeizuführen und nachweislich besser ist als herkömmliche Verfahren. In der bisherigen SA-Forschung in Dramentexten und literarischen Texten allgemein ist eine Ad-Hoc-Verwendung eines beliebigen SA-Lexikons ohne Verwendung besonderer oder auch herkömmlicher Anpassungen in der Verarbeitungskette oder im SA-Lexikon vorherrschend (Nachweis). In der vorliegenden Arbeit soll das SA-Verfahren jedoch, im Gegensatz zur bisherigen Forschung, durch Identifikation und Implementierung von Optimierungsmöglichkeiten nachweislich verbessert werden.

Diesbezüglich wird in der bisherigen Forschung auch keine systematische Evaluation der tatsächlichen Leistung des SA-Verfahrens durchgeführt. Man beschränkt sich meist auf die Bestätigung ausgewählter bekannter Interpretationen. Dieses Vorgehen als alleinige Evaluationsmethodik zu verwenden ist kritisch zu betrachten, da man dazu neigt Interpretationen zu wählen die die eigene SA bestätigen und mögliche Differenzen und Probleme außer Acht gelassen werden. Das vorliegende Projekt will sich durch die systematische Evaluation von Verfahren über verschieden Evaluationsansätze von der bisherigen Forschung absetzen. Dadurch sollen zum einen die gewählten SA-Verfahren miteinander verglichen werden, um den nachweislich besten Ansatz zu identifizieren und zum anderen der grundsätzliche Leistung der SA kontrolliert werden. Die Ergebnisse der systematischen Evaluation tragen zur grundsätzlichen Diskussion des Nutzens von SA für die Literaturwissenschaft bei und können als Benchmark für die Vergleichbarkeit von zukünftigen ähnlichen Projekten dienen.

Die Evaluation findet dabei zweiteilig statt. Als erster Evaluationsansatz wird ein wortbasierter Abgleich der Wörter eines SA-Lexikons mit dem Vokabular des Korpus durchgeführt, also überprüft wie hoch der Anteil an SBWs eines Lexikons im Gesamtkorpus ist. Die genaue Motivation und Ausarbeitung dieses Evaluationsschrittes wird in Kapitel XX beschrieben. Es soll dadurch ein erster grundlegender Eindruck und Vergleich der SA-Lexika erlangt werden. Dieser Evaluationsansatz wird im Folgenden als Vokabular-basierte Evaluation bezeichnet. Als zentraler Evaluationsansatz wird jedoch der Standardevaluationsansatz in der Literatur gewählt, nämlich der Abgleich der eigenen SA mit einem von Menschen annotierten Test-Korpus (Gold Standard; Nachweis). Es wird in der vorliegenden Arbeit nach Kenntnisstand des Autors zum ersten Mal in der Forschung ein mit Sentiment-Informationen annotiertes Dramen-Repliken-Korpus erstellt, um die verschiedenen SA-Verfahren dagegen zu evaluieren. Der Mangel an annotierten Korpora für die SA auf literarischen Texten ist ein zentrales Problem für die Evaluation und die Implementierung fortgeschrittener Methoden. Mit der Erstellung eines annotierten Korpus soll ein erster Beitrag geleistet werden, um diesen Mangel auszugleichen und bezogen auf das eigene Projekt die Auswahl eines optimierten SA-Verfahrens, also des besten SA-Verfahrens, gemäß der gewählten Methoden, ermöglicht werden. Ferner können die entwickelten Evaluations-Frameworks sowohl für die Vokabular-basierte Evaluation als auch die Gold-Standard-Evaluation in zukünftigen ähnlichen Projekten leicht weiterverwendet werden. So lässt sich beispielsweise der Gold-Standard leicht anpassen oder es kann leicht ein anderes Korpus für die Vokabular-basierte Evaluation gewählt werden.

Auf diese Weise werden auch nochmal die Probleme und Herausforderung der Erstellung annotierter Korpora literarischer Texte exploriert (analog zu Alm & Sproat, 2005). Durch die detaillierte Analyse des Annotationsverhaltens und der Annotationsverteilungen kann ein Mehrwert für dieses Forschungsgebiet geliefert werden. Ergebnisse und Erkenntnisse sind auch für die zukünftige Erstellung von Dramen-Korpora bedeutend, die möglicherweise weiterführende Methoden wie Crowdsourcing integrieren. Ferner kann das Annotationsverhalten Aufschluss über die Sentiment-bedingte Konstitution des Gesamtkorpus liefern, z.B. welche Sentiments in welchem Ausmaß vorherrschen und was das bedeutet. Derartige Ergebnisse können zentral für die spätere Interpretation der SA-Daten sein. Insgesamt umfasst die Erstellung der *optimierten SA* also die Entwicklung und Implementierung verschiedener SA-Ansätze, die Erstellung eines annotierten Test-Korpus, die Analyse und Auswertung dieses Korpus und die systematische Evaluation der SA-Ansätze (Vokabular-basiert und gegenüber dem Gold Standard) zur Bestimmung der bestmöglichen Lösung.

Im Folgenden wird nun noch der zweite Teil der Forschungsfrage behandelt. Der *unterstützenden Beantwortung und Diskussion literaturwissenschaftlicher Interpretationen* will sich das vorliegende Projekt durch verschiedene Arbeitsschritte annähern. Es werden SA-Metriken mathematisch definiert und im Kontext möglicher Interpretationen beschrieben, also welche Bedeutung diese in ihrer jeweiligen Ausprägung für die Interpretation haben. Dabei wird auf Erkenntnisse aus der bisherigen Forschung zurückgegriffen und erläutert wie die Metriken eine Hilfe für die Interpretation in der Dramenanalyse darstellen können. Die Berechnung dieser Metriken wird in einem Back-End innerhalb der in den Abschnitten vorher erwähnten Implementierung der optimierten SA-Verfahren integriert. Um den Nutzen und die Verwendungsweise dieser Metriken zu diskutieren wird ein Front-End zur Visualisierung der Metriken der SA entwickelt. Der Mehrwert des Front-Ends ist dabei vielseitig. Zum einen können Möglichkeiten, Grenzen und Nutzen der SA-Visualisierung für die Dramenanalyse exploriert werden. Zum anderen kann eine informelle Überprüfung der Tauglichkeit der SA jenseits der reinen Repliken-Ebene der zuvor beschriebenen Gold-Standard-Evaluation erfolgen. Dazu werden vereinzelte Fallbeispiele aus literaturwissenschaftlichen Interpretationen bezüglich des gewählten Korpus betrachtet. Über die entwickelten Metriken und Visualisierung wird der Nutzen und Mehrwert der SA diskutiert und aufgezeigt und das vielfältige Potential aber auch die Grenzen der SA in der Dramenanalyse erörtert. Dabei handelt es sich um keine systematische Evaluation und Überprüfung, die jedoch mit dem entwickelten Front-End in zukünftigen Projekten erfolgen kann. Ferner kann das Front-End als Anwendung eingesetzt werden um in Zukunft konkrete Anforderungen von Literaturwissenschaftlern zu sammeln und mit diesen an einem konkreten Beispiel Optionen und Wünsche aus Sicht der Geisteswissenschaftler besser zu diskutieren. Insgesamt sollen also zur Beschäftigung mit dem zweiten Teil der Forschungsfrage auf Basis der *optimierten SA* gewinnbringende SA-Metriken speziell für die Dramenanalyse entwickelt werden, Visualisierungen und ein UI konzipiert werden, ein Front-End implementiert werden und vereinzelte Fallbeispiele analysiert und interpretiert werden.

Folgende Tabelle fasst nochmal alle wesentlichen Punkte der Forschungsagenda zusammen. Es werden ferner die Kapitel angegeben, die sich mit der konkreten Ausarbeitung und Lösung der Arbeitsschritte befassen. In den jeweiligen Kapitel werden die einzelnen Schritte im Hinblick ihres Nutzens und im Kontex der Forschung noch mal explizit diskutiert. Des Weiteren werden, falls vorhanden, die jeweiligen produzierten Endprodukte der einzelnen Schritte angegeben.

|  |  |  |
| --- | --- | --- |
| Arbeitsschritt | End- oder Zwischenprodukt | Kapitel |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Obschon die Forschungsfrage also nicht in seiner Gänze beantwortet werden kann, werden durch die aufgestellte Agenda notwendige erste Schritte zur Annäherung und Lösung formuliert und im vorliegenden Projekt auch durchgeführt. Aspekte der Forschungsfrage werden auf unterschiedlichen Ebenen aufgegriffen und ein wesentlicher Mehrwert für den bisherigen Forschungsstand geliefert. Im Bereich der SA in der Dramenanalyse werden zahreiche notwendige Schritte, nach Kenntnisstand des Autors zum ersten Mal durchgeführt: die systematische Entwicklung und Evaluation verschiedener optimierter SA-Verfahren und die Erstellung eines annotierten Sentiment-Korpus auf Replikenebene. Im deutsprachigen Bereich ist kein Projekt bekannt, dass Schritte der Arbeitsagenda bereits umgesetzt hat. Es werden konkrete Probleme und Schwächen der bisherigen Forschung aufgegriffen und weiterverwendbare Endprodukte kreiert. Die Ergebnisse der verschiedenen Bereiche können von zukünftigen Studien aufgegriffen werden um die Forschungsfrage weiter zu diskutieren und die SA in der Literaturwissenschaft unter anderen Gesichtspunkten zu untersuchen.

Zentrale Anknüpfungspunkte werden in Kapitel XX diskutiert. Eine Zusammenfassung der in diesem Projekt erbrachten Leistungen und der Mehrwert und die Implikationen werden auch in Kapitel XX weiter besprochen. Ferner werden die Grenzen und Schwächen der gewählten Agende in Kapitel XX erläutert.

# Dramen-Korpus

Als Untersuchungsgegenstand für die verschiedenen Schritte der Forschungsagenda wurde ein auf einen Autor beschränktes Dramen-Korpus akquiriert. Im Folgenden Abschnitt wird die grundlegende Motivation für diesen Schritt erläutert und einige relevante Korpus-bezogenen Aspekte, die für das Verständnis und die Interpretation einiger Schritte der Forschungsagenda notwendig sind erörtert.

Man hat sich für eine Dramensammlung aus 11 Dramen von Gotthold Ephraim Lessing (1729 – 1781) entschieden. Hauptsächliche Motivation ist ein explizit geäußerter Bedarf einer extern am Projekt beteiligten Literaturwissenschaftlerin. Die Beschränkung ist jedoch auch im Kontext des momentanen Forschungsstandes hilfreich, da noch erste Ergebnisse bezüglich Nutzen und Einsatz von SA gesammelt werden müssen, und größere Vergleiche von verschiedenen Schriftstellern oder Dramengattungen als fortgeschritten betrachtet werden. Ein weitere nützliche Simplifizierung der Beschränkung kommt durch die homogenen Sprache und die leichteren dramenspezifische Vergleichbarkeit, aufgrund desselben Schriftstellers und der gleichen Epoche, zu Stande. So müssen nicht im Größeren Ausmaß stil- und sprachspezifische Besonderheiten bei der Interpretation von Ergebnissen betrachtet werden. Ferner enthält das Dramenkorpus eine angemessen hohe strukturelle und inhaltliche Varianz, also längere und kürzere Dramen, Dramen mit weniger und mehr Figuren, Dramen aus dem Komödienbereich und Tragödien. Auf diese Weise können die entwickelten SA-Metriken und Visualisierungen an ausreichend vielen und unterschiedlichen Fallbeispielen untersucht werden.

Die Beschränkung auf Lessing ermöglicht auch einen klaren Fokus in der Ausarbeitung des Projekts und in der Analyse späterer Fallstudien. Ferner wird auch ein notwendiger begrenzter Arbeitsrahmen gesetzt da innerhalb des Projekts (als auch vergangener Projekte) einzelne Dramen zur Funktionalität häufig angepasst werden mussten und die Anpassung verschiedener Dramen unterschiedlicher Schriftsteller ein nennenswerter Arbeitsschritt zur Integration in die entwickelten Programme ist. Insgesamt ist der Fokus auf einen einzelnen Autor auch momentaner Standard in der Forschung, in der Dramenanalyse wird meist lediglich Shakespeare betrachtet (Mohammad, 2011; Nalisnick & Baird, 2013) aber auch in anderen Bereichen ist eine Reduktion auf einen überschaubaren Korpus eines einzelnen Autors geläufig (z.B. auf Kafka bei Klinger et al., 2016). Die bestehenden Herausforderungen und Probleme der SA werden in der Forschung bereits für kleinere Korpora als ausreichend erachtet, eine weitere Ausbreitung des Korpus wird aus den oben genannten Gründen vermieden.

Die Integration weiterer Dramen liefert zum jetzigen Stand der Forschung keinen relevanten Mehrwert. Die Programme wurden jedoch vereinzelt auf andere Dramen der gleichen Ressource getestet und weisen grundsätzliche Funktionalität aus. Zukünftige Projekte können also leicht weitere Dramen integrieren und damit tiefergehende komplexe Analysen anstoßen.

Die Dramen wurden von der Plattform Textgrid bezogen. Textgrid enthält zahlreiche bekannte deutschsprachige Dramen als XML-Dateien die mit Strukturinformationen ausgezeichnet sind. So werden Metadaten zum Drama (z.B. Autor und Jahr) angegeben und das Drama strukturiert über entsprechende Tags in Akten und Szenen gegliedert. Auf der kleinsten Ebene des Dramas, der Replik wird meist der Sprecher mit seinem Namen angegeben und dann eben der textuelle Inhalt der Replik. Durch Analyse des Schemas können Parser gebaut werden, die den Inhalt des Dramas akquirieren und strukturell angepasst für eine zielspezifische Weiterverwendung speichern. Dies wurde bereits im Vorgängerprojekt Katharsis durchgeführt. In der vorliegenden wurde der Parser jedoch noch erweitert und für den SA-Einsatz angepasst sowie kleinere Fehler behoben (siehe auch Kapitel XX). Folgender Screenshot illustriert für einen Teil die Art und Weise der XML-Auszeichnung in den TextGrid-Dramen:

Es wird an dieser Stelle die Art der XML-Auszeichnung nicht im Detail erläutert. Es ist jedoch festzuhalten, dass die Dramen nicht stets einheitlich und korrekt annotiert sind, insbesondere auf der Repliken-Ebene. Der entwickelte und angepasste Parser ist hinsichtlich des gewählten Korpus optimiert und verarbeitet im Wesentlichen etwa drei grundsätzliche Annotationsschema (von denen manchmal auch mehrere in einem Drama erscheinen können). Ferner mussten die Annotationen einiger Dramen zu korrekten Verarbeitung noch angepasst werden. In Nathan der Weise wird beispielsweise aufgrund einer fehlerhaften Annotation eine Figur gar nicht erfasst, dies musste manuell ausgebessert werden. Derartige und einige kleinere andere Probleme manifestierten sich an unterschiedlichen Stellen des Projekts und wurden dann händisch angepasst. Aus Modellierungsgründen wurden für kleinere Dramen, die nur aus Szenen bestehen, eine Akt-Annotation eingebaut, so dass diese nun aus einem Akt bestehen. Im Anhang sind sowohl die Roh-Dramen enthalten als auch die für das Projekt angepassten und ausgebesserten XML-Dramen-Dateien.

Zum tieferen Verständnis des Projekts wird für das Korpus noch der literarische und historische Kontext knapp skizziert. Lessing wird als bedeutender Dichter der deutschen Aufklärung betrachtet, der die Entwicklung des Theaters nachhaltig beeinflusst hat und der als erster deutscher Dramatiker gilt, dessen Werke bis heute ununterbrochen in Theatern aufgeführt werden (Gotthold Ephraim Lessing, o. J.). Er konnte sich Zeit seines Wirkens einer hohen Wertschätzung in der Rezeption erfreuen (Fick, 2016; S. 1). Das hier gewählte Dramenwerk lässt sich grob in zwei Gruppen zergliedern. Damon oder die wahre Freundschaft (1747), Der Misogyn (1748), Der junge Gelehrte (1748), Die alte Jungfer (1748), Die Juden (1749), Der Freigeist (1749) und Der Schatz (1750) entstanden zwischen 1747 und 1750 in Leipzig (Pelster, 2017, S. 94-95). Es handelt sich bei allen um Lustspiele, also vereinfacht gesagt um Komödien (Lustspiel, o. J.). Die Dramen werden in der Literaturwissenschaft ot als Vorarbeiten und Übungen betrachtet (Pelster, 2017, S. 95). Lessing setzte sich für diese Dramen zum Ziel der deutsche Molière (bekannter französischer Komödiendramatiker) zu werden (Pelster, 2017, S. 105).

Die zweite Gruppe bilden die bekanntesten Dramen in Lessings Werk: Miss Sara Sampson (1755, Bürgerliches Trauerspiel), Philotas (1759, Trauerspiel) Minna von Barnhelm oder Das Soldatenglück (1767, Lustspiel), Emilia Galotti (1772, Bürgerliches Trauerspiel) und Nathan der Weise (1779, Dramatisches Gedicht, 1779). Bei der zweiten Gruppe handelt es sich also gemäß dieser Differenzierung um das Spätwerk (1755-1779). Es entstanden in der Zeit auch Trauerspiele, also wieder vereinfacht betrachtet Tragödien. Der Begriff des bürgerlichen Trauerspiels grenzt sich von der herkömmlichen Tragödie vor allem dadurch ab, dass nicht mehr Adelige im Mittelpunkt stehen, sondern Angehörige des bürgerlichen Standes (Pelster, 2017, S. 119). Miss Sara Sampson gilt als der Prototyp des bürgerlichen Trauerspiels (Pelster, 2017, S. 96).

Insgesamt wird Lessing in seinen Dramen eine prosaische natürliche Sprache attestiert (Gotthold Ephraim Lessing, o. J.), die sich vor allem in der Nicht-Verwendung von Versformen verdeutlicht. Einzige Ausnahme bildet das Drama Nathan der Weise, dass im Blankvers geschrieben ist (Nathan der Weise, o. J.). Dies äußert sich auch in einer abweichenden XML-Annotation für das TextGrid-Drama. Es wird angenommen, dass die natürliche prosaische Sprache von Vorteil für die im Projekt verwendete SA-Methodik mittels Lexika, da die SA-Lexika aufgrund ihres aktuelleren alltäglicheren Wortschatzes zusätzlich problematisch auf die poetische und speziellere Sprache von Dramen in Versform reagieren. Diese Aussage stellt keine belegte Feststellung dar, sondern nur eine informelle Annahme, die in zukünftigen Projekten genauer betrachtet werden kann. Für das bestehende Korpus wird jedoch aufgrund der genannten Annahme explizit auf das Drama Nathan der Weise geachtet. Die grundsätzlichen Probleme aufgrund der poetischen Sprache literarischer Texte und aufgrund der großen zeitlichen Abstands bleiben bestehen. Tiefergehende Informationen zu Leben und Werk Lessings sowie zur Interpretationshistorie findet man bei Fick (2016).

Folgende Tabelle listet alle Dramen des Korpus auf und gibt relevante Meta-Daten, Repliken und Wort-Statistiken an.

#TODO Tabelle und Beschreibung

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Titel | Jahr | Gattung | Zahl der Repliken | Prozentualer Anteil | Längste Replik | Replikenlänge (Avg) | Anzahl an Wörtern |
| Damon oder die wahre Freundschaft | 1747 | Lustspiel |  |  |  |  |  |
| Der Freigeist | 1749 | Lustspiel |  |  |  |  |  |
| Der junge Gelehrte | 1748 | Lustspiel |  |  |  |  |  |
| Der Misogyn | 1748 | Lustspiel |  |  |  |  |  |
| Der Schatz | 1750 | Lustspiel |  |  |  |  |  |
| Die alte Jungfer | 1748 | Lustspiel |  |  |  |  |  |
| Die Juden | 1749 | Lustspiel |  |  |  |  |  |
| Emilia Galotti | 1772 | Bürgerliches Trauerspiel |  |  |  |  |  |
| Minna von Barnhelm oder Das Soldatenglück | 1767 | Lustspiel |  |  |  |  |  |
| Miss Sara Sampson | 1755 | Bürgerliches Trauerspiel |  |  |  |  |  |
| Nathan der Weise | 1779 | Dramatisches Gedicht |  |  |  |  |  |
| Philotas | 1759 | Trauerspiel |  |  |  |  |  |

Die nachfolgende nun fasst die relevanten Daten für das Gesamtkorpus zusammen:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |

Die einzelnen Statistiken sowohl pro Drama also auch gesamt sind sehr relevant für die Interpretation der Ergebnisse einzelner Projektbestandteile und später auch zur Erstellung eines adäquaten Evaluations-Test-Korpus.

# Back End: Sentiment Analysis

## Konzeption

## Vorverarbeitung der Dramen

### Sprachverarbeitung – Lemmatisierung

#### Idee

#### Entwicklung

## Lexika-Verarbeitung

### Grundsätzliche Verarbeitung

### Lexika-Kombination

### DTA – Erweiterung

#### Idee

#### Entwicklung

### Sprachverarbeitung - Lemmatisierung

#### Idee

#### Entwicklung

### Stoppwortlisten

#### Idee

#### Entwicklung

## Vorverarbeitung – Überblick

## Sentiment Analysis

### Konzeption

### Entwicklung

### SA-Metriken

# Lexika-Analyse

## Idee

## Entwicklung

## Ergebnisse

## Diskussion

# Vokabular-basierte Evaluation

## Idee

## Entwicklung

## Ergebnisse

## Diskussion

# Erstellung des Gold-Standard-Korpus

Die Erstellung des eines mit Sentiment-Informationen annotierten Gold-Standard-Korpus (Test-Korpus) ist ein notwendiger Schritt zur systematischen Evaluation der SA-Verfahren. Dabei wird anhand der von Menschen angegebenen Annotation für die einzelnen Korpus-Einheiten überprüft ob und in welchem Ausmaß die Ausgabe der SA mit den Annotationen übereinstimmt. Je höher diese Übereinstimmung als desto besser wird die SA angesehen. Es handelt sich um das Standard-Verfahren zur Evaluation in der SA (Vinodhini & Chandrasekaran, 2012). Die Annotationen können dabei bereits implizit vorliegen, z.B. bei der Nutzung von Stern-Wertungen bei Produkt- (Pang & Lee, 2004) oder Film-Reviews (Pang, Lee & Vaithyanathan, 2002) als Angaben für die Polarität oder durch explizite Empfehlungsangaben von Autoren in Reviews (Turney, 2002). Kouloumpis et al. (2011) verwenden die Polarität von Twitter-Hashtags als implizite Annotation für Tweets. Ferner werden aber auch Polaritätsangaben durch manuelle Annotationen von Menschen akquiriert (z.B. Bosco et al., 2014; Refaee & Rieser, 2014; Mozetic et al., 2016). Die letztgenannte Methodik wird trivialerweise meist dann durchgeführt wenn kein automatisch annotiertes Korpus vorhanden ist, also bei Daten wie Blogs (Kessler et al., 2010) oder bei sehr speziellen Domänen, bei der für die Korpus-Erstellung Expertenwissen notwendig ist, wie z.B. News zum Finanzwesen (Malo et al., 2013, Takala et al., 2014). Oft wird ein annotiertes Test-Korpus sowohl für das Training eines ML-Algorithmus als auch für dessen Evaluation genutzt (Wilson, Wiebe & Hoffmann, 2005). Für zahlreiche gängige Anwendungsgebiete der SA gibt es bereits standardisierte Gold-Standards die zum projektübergreifenden Vergleich von SA-Methoden verwendet werden können. Eine grobe Übersicht für derartige Gold-Standard-Ressourcen verschiedener Anwendungsfelder findet man bei Takala et al. (2014) und Tsytsarau und Palpanas (2012).

Insbesondere im Bereich literarischer Texte werden bislang eher informelle Evaluationen durch anekdotische Überprüfungen von bekannten literaturwissenschaftlichen Interpretationen durchgeführt (Mohammad, 2011; Nalisnick & Baird, 2013). Die vorliegende Studie will diesen Mangel in der Forschung ausgleichen und orientiert am sonstigen SA-Gebiet die SA-Verfahren auch systematisch und objektiv evaluieren. Das Test-Korpus kann als erster grundsätzlicher Gold-Standard in zukünftigen Studien genutzt werden, aber auch noch erweitert und optimiert werden. Ferner werden Probleme und Herausforderungen der Annotation literarischer Texte, analog zu Alm und Sproat (2005) nun aber speziell für das deutschsprachige Drama des 18. Jahrhunderts, exploriert.

Die Erstellung des Gold-Standard gliedert sich in drei größere Bereiche. Die semi-automatische Erstellung des Korpus an sich, also die Auswahl der Repliken, die Planung und Durchführung der Annotation sowie die Auswertung der Annotation und des Annotationsverhalten. Abschließend werden die Ergebnisse diskutiert.

## Test-Korpus-Erstellung

### Idee und Vorgehen

Die Erstellung des Test-Korpus orientiert sich am grundsätzlichen Verfahren aus der Forschung dazu und an den Möglichkeiten, die aufgrund des Datensatzes zur Verfügung stehen. Das Korpus enthält auf keiner Ebene explizite Sentiment-Annotationen. Als implizite Annotation kann die Gattungszugehörigkeit eines einzelnen Dramas angenommen werden. Dies führt jedoch zu einer zu starken Generalisierung und dadurch dass es sich um die höchste strukturelle Ebene handelt verhindert dies differenzierte Analysen. Ferner wurde auch die Annotation von anderen hochstufigen Ebenen, wie der Akt-, Szenen- oder Sprecher-Ebene abgelehnt. Manuelle Annotation dieser Ebenen erfordert großes Expertenwissen. Im Rahmen der vorliegenden Arbeit bestand kein Zugriff auf zahlreiche domänenspezifische Experten. Ferner wäre eine diesbezüglich Annotation aufwendig und komplex. Die kleinsten Ebenen der Dramenanalyse sind nun die Replik und der Satz. Da die Annotation von Sätzen als speziell schwer angesehen wird (Nachweis) und die SA auf Sätzen ebenso als besonders herausfordernd betrachtet wird, hat man sich für die vorliegende Arbeit für die Replik Einheit für die Annotation entschieden. Die Replik ist die zentrale und kleinste strukturelle Einheit eines Dramas. Über die Repliken ist es möglich einen breiten Ausschnitt des Gesamtkorpus zu betrachten und verschiedene linguistische und strukturelle Besonderheiten differenziert zu betrachten. Ferner gibt es auch rein praktische Gründe. Die Replik ist aufgrund ihrer angemessenen Kürze besser geeignet für eine manuelle Annotation als größere Abschnitte (wie Szenen). In der Diskussion in Kapitel XX werden jedoch auch einige Problem dieses Fokus auf die Replik erläutert.

Mit der Konstitution des Test-Korpus wird angestrebt ein adäquates Abbild des Gesamt-Korpus zu erhalten aber gleichzeitig notwendige Bedingungen für die SA und die Annotation zu erfüllen. Das Gesamtkorpus besteht aus insgesamt 8224 Repliken unterschiedlicher Länge (siehe Kapitel XX). In der Forschung werden meist etwa 10% des Gesamtkorpus als Testkorpus gewählt (Nachweis). Dies entspräche etwa 800 Repliken. Es ist abzuschätzen, dass die Annotation von 800 Repliken sehr anspruchsvoll und zeitaufwändig verläuft. Die bisherigen Ergebnisse zur Annotation von literarischen Texten legen dies nahe (Alm & Sproat, 2005; Nachweis). Im Rahmen des Projekts bestand ferner keine Möglichkeit einen monetären Inzentiv zu bieten. Um also die Annotation angemessen zu halten und mögliche Übermüdungserscheinungen und damit einhergehende fehlerhafte Annotationen zu vermeiden hat man sich für die Erstellung eines deutlich kleineres Korpus von 200 Repliken entschieden. Dies entspricht lediglich etwa 2% des Gesamtkorpus. Dabei muss man jedoch bedenken, dass ein großer Teil des Korpus aus sehr kurzen Repliken besteht. Dennoch muss die begrenzte Größe für die zukünftige Interpretation mitbetrachtet werden. Bei Projekten zur SA auf literarischen Texten, die in irgendeiner Weise ein Test-Korpus erstellen ist ebenso erkennbar, dass aufgrund des Annotationsaufwands vergleichsweise kleine Korpora erstellt werden (Marchetti et al., 2016).

Die Repliken-Statistiken zeigen, dass die Wortlänge der Repliken einen Median von 13 und einen Mittelwert von 24 aufweisen. Das heißt das Korpus enthält sehr viele, sehr kurze Repliken. Sowohl bei für Annotatoren als auch für die lexikonbasierte SA stellen kurze Sätze/Repliken größere Probleme dar (Nachweis). Annotatoren als auch die SA haben dann weniger Informationen zur Verfügung und die Entscheidung fällt in einem größerem Maße zufällig aus. Aus diesen Gründen hat man sich dazu entschieden die Länge der Repliken für das Testkorpus insofern zu kontrollieren, dass nur Repliken ab einer bestimmten Länge aufgenommen werden. Zum jetzigen Stand der Forschung wird dieser Schritt als legitim betrachtet um größere Probleme bei der Annotation und der SA zu umgehen und sich auf die Produktion brauchbarer Ergebnisse zu fokussieren. Um ausreichend lange Repliken für eine angemessen Interpretation zu erlangen hat man sich entschieden nur Repliken in den Korpus aufzunehmen die mindestens 19 Wörter enthalten. Dies entspricht in etwa mindestens ein Wort mehr als -25% des Durchschnittswertes der Replikenlänge (25% von 24 = 6; 24-6=18). Durch Analyse einiger zufällig erstellter Korpora erzielte diese Heuristik zufriedenstellende Ergebnisse insofern, dass Repliken der Wortlänge 19 angemessen für eine korrekte Annotation und Interpretation sind, während kleinere Repliken als problematisch betrachtet werden und meist auch kaum Informationen enthalten. Andere Heuristiken, die noch stärker längere Repliken einbeziehen (z.B. nur >24) entfernen sich zu stark vom Prinzip dem Gesamtkorpus zu ähneln, da kurze 1-2-Satz-Repliken einen großen Teil des Korpus ausmachen. Generell können zukünftige Studien das Längenkriterium weglassen, um sich mit den Herausforderungen von sehr kurzen Repliken sowohl auf Annotations- als auch SA-Ebene zu befassen.

Als weiteres Kriterium musste eine Replik sowohl eine Vorgänger- als auch eine Nachfolger-Replik enthalten. Mit Hilfe dieses Kontext soll die korrekte Annotation für die Annotatoren erleichtert werden und der Dialog bezüglich einer Replik in die Annotationsentscheidungen miteinfließen. Dabei orientiert man sich an der Idee von Alm und Sproat (2005 Nachweis), die ebenfalls für Sätze von Märchen stets den vorigen Satz und den nachfolgenden Satz mitangaben, um dem Annotator inhaltlichen Kontext zu ermöglichen. Speziell für den vorliegenden Fall hat man sich dafür entschieden Repliken die am Anfang oder Ende eines Akts oder einer Szene stehen nicht für den Test-Korpus zu erlauben da keine direkten Vorgänger- oder Nachfolger-Repliken für diese Repliken vorliegen. Diese befinden sich in der vorigen Szene oder im vorigen Akt. Dies kann bei der Annotation zu Problemen bei der Interpretation des Inhalts führen und unnötige Verwirrung stiften. Häufig sind zwar szenenübergreifende Repliken durchaus verständlich und aufeinander bezogen, dennoch wurde aufgrund der potentiellen genannten Probleme dieses Kriterium eingehalten. Bezüglich des Kontextes muss man jedoch kritisch anmerken, dass dieser durch die Vorgänger- und Nachfolger-Replik nicht vollständig hergestellt werden kann. Zum genauen Verständnis einer Replik ist größeres Wissen über die Gesamthandlung und das ganze Drama notwendig. Diese konnten in der vorliegenden Umsetzung nicht mitgeliefert werden. Man kann auch behaupten, dass nur Leser und Kenner des Werkes eine Replik in einen Gesamtkontext korrekt einordnen können, was von den hier gewählten Annotatoren nicht erwartet wurde.

Des Weiteren wurde um eine passendes Abbild des Gesamtkorpus darauf geachtet, dass der verhältnismäßige Replikenanteil einzelner Dramen am Gesamtkorpus auch im Test-Korpus eingehalten wird. Längere Dramen, gemäß Replikenzahl, sind also mit mehr Repliken vertreten und kürzere Dramen mit weniger. Enthält ein Drama beispielsweise 10% der Repliken des Gesamtkorpus enthält es auch 10% im Testkorpus (also etwa 20 Repliken). Die Verhältnismäßigkeit geht dabei nicht exakt auf, kann jedoch durch vereinzelte Rundungen fast vollständig erreicht werden. Über diese Aufteilungen können später auch Ergebnisse pro Drama betrachtet werden, die für Dramen die häufiger mit Repliken im Test-Korpus vertreten sind aussagekräftiger sind.

Gemäß aller genannter Kriterien, Mindestlänge, das Vorhandensein einer Vorgänger- und Nachfolger-Replik sowie die Verhältnismäßigkeit pro Drama wurden ansonsten zufällig akquiriert und zusammengestellt (mittels der in Kapitel XX beschriebenen Programme). Es wurden mehrere Test-Korpora erstellt und bezüglich verschiedener Replikenlängen (Median, Mittelwert, Maxium, Minimum) und der grundsätzlichen Längenverteilung analysiert. Es wurde dann das Test-Korpus, der sich bezüglich Metriken am ausgeglichensten präsentierte gewählt.

Ferner wurde das Test-Korpus noch manuell kontrolliert auf besondere Fälle. So wurden vereinzelte Repliken, die beispielsweise Französische Sprache enthielten oder zu viele altertümliche Worte enthielten ausgetauscht. Das finale Korpus wird in Kapitel XX besprochen.

#genaue Zahlenwerte >19 usw.

# Größe Nachweise

# Annotationsoptionen Nachweise, Annotatoren

# Skalen Nachweise

# Vorgänger und Nachfolger Sätze

# Metriken, Fleiss Kappa, Krippendorffs Alpha, Prozentuale Übereinstimmung, Mehrheiten, Mehrheitswahl

### Entwicklung

Zur Erstellung des Test-Korpus wurden ein Python-Programme entwickelt:

evaluation\_test\_corpus\_creation.py

Das Programm erlaubt über die Nutzung der Klasse Test\_Corpus\_Creator die Kreation von zufällig zusammengestellten Test-Korpora nach den oben beschriebenen Kriterien. Das Objekt enthält Attribute zur Anpassung der Größe des Korpus und auch zur Einstellung der Anzahl an Repliken pro spezifischem Drama (\_testCorpusSizeFactor, \_partsPerDrama). Die Klasse enthält ferner Methoden zur Erstellung eines derartigen Korpus, zur Anpassung des Korpus, zum Austausch einzelner Repliken oder Replikengruppen eines bestehenden Korpus, und zur Ausgabe und Abspeicherung eines Test-Korpus als txt-Datei oder Pickle-Datei (also als mit Python verarbeitbare Datei).

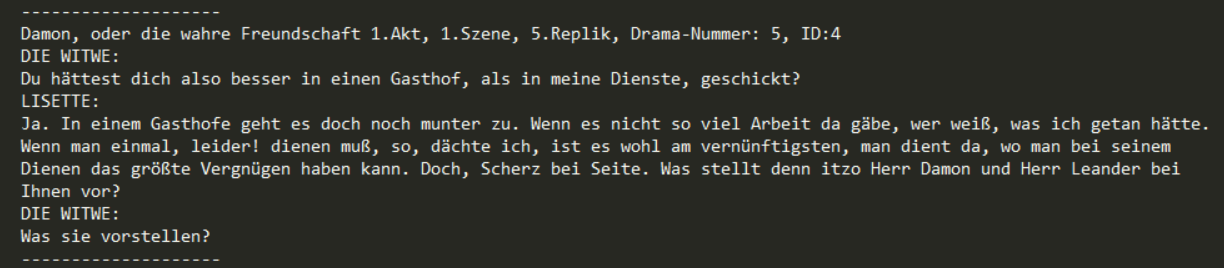
Die Klasse greift dabei auf Objekte der Klasse Test\_Corpus\_Speech zu. Dabei handelt es sich um eine speziell für die Aufgabe der Test-Korpus-Erstellung kreierte Klasse zur Modellierung von Repliken im Test-Korpus. Es handelt sich dabei um eine ausgebaute Form der herkömmlichen Replik aus dem erweiterten Dramen-Modell. So besteht Test\_Corpus\_Speech neben der Original-Replik auch noch aus notwendigen Meta-Informationen für die Korpus-Erstellung und der Vorgänger- und Nachfolger-Replik. Die jeweiligen Attribute werden während der Test-Korpus-Erstellung im Test-Corpus-Creator gesetzt.

Die Klasse Test\_Corpus\_Handler kann einen Test-Korpus als Pickle-Datei einlesen und zur Analyse verarbeiten. Auf diese Weise wurden beispielsweise Lage- und Verteilungsmaße bezüglich der Replikenlängenverteilung betrachtet um einen optimalen Test-Korpus zu finden.

### Ergebnisse

Das komplette Test-Korpus ist im Anhang einsehbar als txt-Datei, als Word-Dokument (wie es auch die Annotatoren erhalten haben) und als Pickle-Datei zur Analyse und Weiterverwendung in Python als Liste von Test\_Corpus\_Speech-Objekten im Anhang einsehbar.

Eine Test-Korpus-Einheit besteht dabei zunächst aus einer Metazeile, die das Drama, die strukturelle Position im Drama sowie die ID bezogen auf das ganze Test-Korpus angibt. Als nächstes folgend die drei Repliken, von denen die Annotations- und Bezugs-Replik die mittlere ist. Davor wird die Vorgänger-Replik und danach die Nachfolger-Replik angegeben. Für beide wird auch der Sprecher vorgestellt. Folgender Screenshot zeigt eine Beispiel-Replik aus der txt-Datei:



Folgende Tabelle beschreibt die wesentlichen Statistiken des Test-Korpus:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

#TODO

## Test-Korpus-Annotation

### Idee und Vorgehen

Zur Erhaltung eines Gold-Standard musste der zuvor beschriebene Test-Korpus auch mit Sentiment-Annotationen ausgezeichnet werden. Bei der Erstellung der Annotationsanweisung, als auch bei der Durchführung hat man sich an ähnlichen Verfahren in der Forschung orientiert.

#### Annotationsschema

Das zentrale in der vorliegenden Arbeit betrachtete Sentiment ist die Polarität, also die Bewertung ob eine Replik als positiv oder als negativ assoziiert wird. Aus diesem Grund wurde besonders auf eine differenzierte und gewinnbringende Erhebung dieses Attributs geachtet. Ziel ist nicht nur die Erstellung eines annotierten Korpus sondern auch die Analyse von Annotationsverhalten und Annotationsergebnissen um Rückschlüsse auf die Sentiment-bedingte Konstitution des Korpus zu machen. Aus diesem Grund muss ein passendes und zielführendes Annotationsschema für die Polarität der Repliken entwickelt werden.

In der Forschung findet man verschiedene Annotationsschemata zur Erfassung der Polarität. Wiebe, Wilson und Cardie (2005) beschäftigen sich ausführlich mit den Problemen und Herausforderungen der Annotation von Meinungen und Gefühlen in Texten. Sie beschreiben ein komplexes Modell aus verschiedenen Bestandteilen von denen vor allen Dingen einige Eigenschaften eines subjektiven Sprachereignisses Einfluss auf spätere Annotationsstudien in der SA gefunden haben. Als Eigenschaften definieren sie die Intensität, die Ausdrucksintensität, die Bedeutsamkeit und den Einstellungstyp. Vor allem die Intensität und der Einstellungstyp werden in Studien und auch in der vorliegenden Arbeit aufgegriffen. Der Einstellungstyp entspricht der Polarität und wird von Wiebe et al. (2005) zwischen positiv, negativ, neutral und anders unterschieden. Die Intensität und die Ausdrucksintensität verläuft 4-stufig von gering bis extrem. Van de Kauter, Desmet und Hoste (2015) bezeichnen das Annotationsschema von Wiebe et al. (2005) als das bekannteste Schema in der SA.

Die Ideen wurden in der SA bei Annotationsstudien aufgegriffen. Als simpelste Form findet man die einfache Angabe von positiv oder negativ für die Polarität (Nachweis). Zur Erhebung differenzierterer Informationen findet man jedoch auch Annotationsschemata mit mehr Auswahlmöglichkeiten, orientiert an Wiebe et al. (2005). Es werden einige Beispiel genannt, die Einfluss auf die Überlegung zur eigenen Erstellung eines Annotationsschemas hatten. Bei Bosco et al. (2014) werden Tweets in den Kategorien positiv, negativ, objektiv, gemischt und unverständlich angegeben. Ein ähnliches Schema wählen Refaee und Rieser (2014) in der Unterscheidung zwischen positiv, negativ, neutral, gemischt und anders/ungewiss. Takala et al. (2014) verlangen von ihren Annotatoren eine Bewertung auf einer sehr differenzierten 7-stufigen Polaritätsskala von sehr negativ bis sehr positiv. Auf diese Weise erhalten Sie im ersten Schritt Daten für die detaillierte Analyse des Annotationsverhaltens und der Korpus-Zusammensetzung. Für die finale Annotation bilden sie die Bewertung anschließend auf einer dreistufigen Skala ab: positiv, neutral, negativ. Momtazi (2012) verwendet eine ähnlich differenzierte Skala, analysiert aber anschließend die binäre Polarität und die Stärke auch getrennt. Ein komplexes Annotationsschema, dass zwischen verschiedenen Typen und Ausprägungen unterscheidet wird von Shin et al. (2012) vorgeschlagen. Zentrale Aspekte sind aber auch hier pro Typ die Unterscheidung zwischen positiv, negativ, neutral und hier komplex (was sich ähnlich zu gemischt verhält).

Im Bereich literarischer Texte wird sich meist nicht auf die Annotation von Polarität beschränkt. Bei Alm und Sproat (2005) werden Sätze bezüglich Vorhandensein von Emotionskategorien und Neutralität beurteilt. In einem anschließenden Schritt differenzieren aber auch sie zwischen positiven und negativen Emotionsgruppen und somit zwischen Polaritäten. Für ihren ML-Ansatz (Alm et al., 2005) unterscheiden sie dann zwischen Emotion vorhanden und Emotion nicht vorhanden. Auch Volkova et al. (2010) gehen bei der Annotation von Märchen emotionsbasiert mit sehr differenzierten Emotionskategorien vor. Bei Volkova et al. (2010) markieren Annotatoren konkrete Textstellen als emotionstragend. Marchetti et al. (2014) beschränken sich wieder auf die Polarität und lassen Sätze ihrer historischer Texte herkömmlich bezüglich Polarität mit den Gruppen positiv, negativ, neutral und unbekannt, bewerten. Alle genannten Studien zu literarischen Texten werden auch in Kapitel XX genauer beschrieben. Ein weiteres komplexes Schema sowie eine Aufarbeitung relevanter Literatur zum Thema Annotation in der SA findet man bei Van der Kauter et al. (2015).

Als zentraler Konsens in der Forschung ist die Verwendung von den Gruppen positiv, negativ, neutral und gemischt vorherrschend. Manchmal wird positiv und negativ noch bezüglich ihrer Intensität unterscheidet (siehe oben). Ferner fällt aber auch auf, dass derartige starke Differenzierungen zu geringen Übereinstimmungen in der Annotation führen was wiederum zu Schwierigkeiten bei der Interpretation und Gold-Standard-Erstellung führt, weswegen dann für die Auswertung häufig auf Oberkategorien zurückgegriffen wird (Nachweis). Auf Basis der bestehenden Lösungen hat man sich für das folgende Schema entschieden. Annotatoren geben zunächst die Polarität einer Replik auf einer nominalen Skala an: sehr negativ, negativ, neutral, gemischt, positiv, sehr positiv. Durch diese Aufgliederung kann man ein differenziertes Bild über die Annotation und die Polaritätsverteilung im Korpus erhalten. Für die tatsächliche Gold-Standard-Erstellen müssen die Teilnehmer jedoch in einem zweiten Schritt explizit angeben, ob eine Replik eher positiv oder negativ ist, also auf einer binären Skala die Polarität angeben. Auf diese Weise werden auch zuvor neutrale oder gemischte Repliken mit Annotation für die Polarität ausgezeichnet.

Folgende Tabelle illustriert das Annotationschema für die Polarität:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sehr negativ | Negativ | Neutral | Gemischt | Positiv | Sehr Positiv |
|  |  |  |  |  |  |
|  |  |  | |  |  |

Der letztgenannte Schritt ist kritisch zu betrachten, aufgrund der geringen Größe des Korpus hat man sich aber dafür entschieden um notwendige Annotation für die spätere Auswertung zu halten. Die damit verbundenen Probleme werden bei der jeweiligen Interpretation jedoch stets beachtet. Grundsätzlich kann man über die Aufteilung aber auch Fragen das Annotationsverhalten genauer betrachtet, z.B. inwieweit Annotatoren sich bezüglich der differenzierten Skala übereinstimmen im Vergleich zur binären.

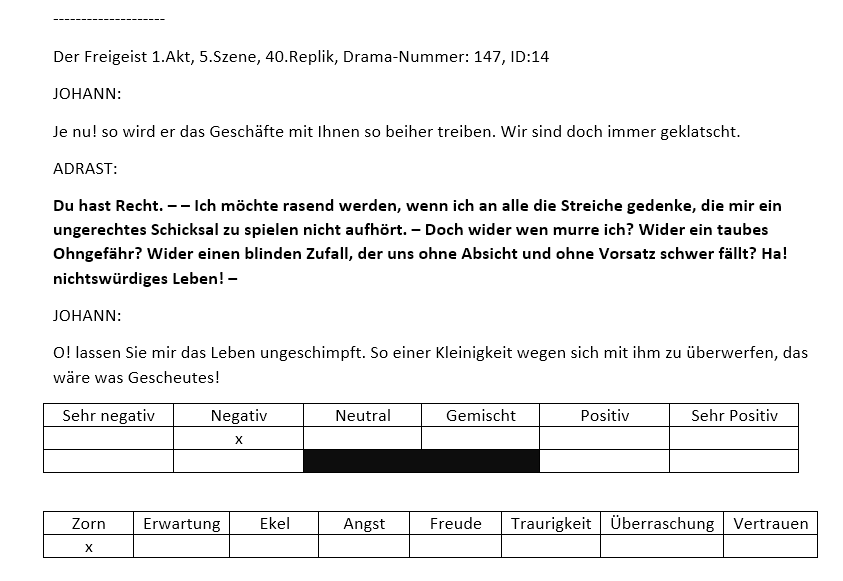
Auf Basis der Literaturanalyse kann man festhalten, dass komplexe Emotionskategorien für literarische Texte besonders bedeutend sind (Nachweis). Obschon in dieser Studie die Polarität als zentrale Sentiment-Kategorie angesehen wird, wurden auch Annotationen zu den Emotionskategorien des NRC eingebaut. Auf diese Weise können erste Erfahrungen zur Emotionsannotation in Dramen gemacht werden und in Beziehung zu den bisherigen Ergebnissen im Bereich der Annotation von Märchen gesetzt werden (Alm & Sproat, 2005). Das Annotationsschema wird durch folgende Tabelle illustriert. Teilnehmer geben lediglich an, ob eine Emotion vorliegt oder nicht:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Zorn | Erwartung | Ekel | Angst | Freude | Traurigkeit | Überraschung | Vertrauen |
|  |  |  |  |  |  |  |  |

#### Durchführung und Stichprobe

Für alle Annotatoren wurde manuell ein Word-Dokument aus den Repliken des Test-Korpus und den obigen Tabellen zur Annotation erstellt. Im Test-Korpus wird dazu erst die Replik angegeben, bestehend aus einer knappen Metazeile, der Vorgänger-Replik, der zu annotierenden Replik und der Nachfolger-Replik. Die zu annotierende Replik ist dabei fett gedruckt. Anschließend folgen die obigen Annotationstabellen. Teilnehmer sollten in jede Tabelle ein Kreuz als X eingeben für diejenige Ausprägung die ihres Erachtens am meisten zu der Replik passt. Also im Fall von Polarität, welche Polarität sie am meisten mit der Replik assoziieren. Zunächst sollten sie dafür eine Annotation für die differenzierte Skala machen, sollten sie dabei keine Negativitäts- oder Polaritätskategorie ankreuzen, sollen sie in der nächsten Zeile wieder über ein X angeben welche Polarität trotzdem am ehesten mit der Replik assoziiert wird. Für die Emotionskategorien konnten die Annotatoren beliebig viele Emotionen als vorhanden markieren, also auch keine. Alle Annotatoren wurden mündlich anhand des Word-Dokuments in den Ablauf und die Durchführung eingeführt. Sie konnten sich die Zeit zur Bearbeitung individuell aufteilen und von zu Hause arbeiten. Aufgrund der erwartbaren kognitiven Herausforderung und Anstrengung wollte man fehlerhafte Annotationen vermeiden und den Annotatoren auf diese Weise genug Zeit zur korrekten Bearbeitung einräumen.

Das Word-Template zur Annotation findet man im Anhang. Hier wird ein komplettes Beispiel für eine Replikenannotation mit ausgefüllten Tabellen zum besseren Verständnis gezeigt:



An der Annotation haben 5 Studenten, allesamt deutsche Muttersprachler teilgenommen, d.h. jede Replik wurde von 5 Personen annotiert. Die Zahl verhält sich konform zu ähnlichen Vorgehen in der Forschung. Da die Aufgabe der Annotation sehr komplex ist und von der Interpretation von Einzelpersonen abhängt wird Sentiment-Annotation meist von mehreren Personen durchgeführt und Übereinstimmungen sowie Annotationsverteilungen analysiert. Als Mindestgröße dient dabei 2. Refaee und Rieser (2014) verwenden beispielsweise die Mindestgröße von 2 Personen zur Annotation, auch Alm und Sproat (2005) lassen die Sätze aus Märchen von 4 Annotatoren in Zweier-Gruppen bewerten. Viele Studien nutzen jedoch eine ungerade Zahl von Annotatoren, z.B. 3 (Momtazi, 2013; Takala et al., 2014; Kohler, Kruge & Gurevych, 2015). Marchetti et al. (2016) verwenden neben 2 Experten-Annotatoren auch Crowdsourcing mit je 5 Annotationen pro Satz. Auch Mohammad und Turney (2010) nutzen Crowdsourcing mit je 5 Annotationen pro Wort. Bosco et al. (2014) reduzieren im letzten Schritt ihre Annotationen von einer ungeraden Zahl auf eine gerade. Der Vorteil von einer ungeraden Anzahl ist, dass für die Gold-Standard-Erstellung bei einer binären Polaritätsangaben Mehrheitsangaben gewählt werden können, d.h. eine Bewertungseinheit erhält diejenige Polarität, die die Mehrheit der Annotatoren auswählt. Bei lediglich zwei Annotationen oder anderen geraden Zahlen können sonst gleichmäßige Bewertungen auftreten und ein künstlicher „Tie-Breaker“ wird notwendig (z.B. bei Alm & Sproat, 2005). Aus diesem Grund wurde auch in der vorliegenden Arbeit eine ungerade Zahl angestrebt. In manchen Bereichen findet man auch größere Zahlen an eingesetzten Annotatoren, z.B. 10 bei Volkova et al. (2010) oder 16 bei Malo et al. (2013). Basierend auf den Möglichkeiten wurde jedoch mit 5 Annotatoren eine angemessene und für die Forschung vergleichbare Anzahl gewählt.

Des Weiteren wurde die Idee eine größere Menge von Repliken von einer kleineren Zahl von Annotatoren auszeichnen zu lassen, wie es in anderen Studien durchgeführt wird (Alm & Sproat, 2005; Volkova et al., 2010), nicht aufgegriffen um das Annotationsverhalten und die Übereinstimmung einer angemessen großen Zahl von Annotaren aussagekräftig für einen begrenztes Korpus zu analysieren.

Ferner sei zu beachten, dass in der Forschung, vor allem bei Spezial-Themen, Teilnehmer häufig eine Experten-Ausbildung besitzen, z.B. bei Takala et al. (2014) bei der Annotation von News aus dem Finanzsektor. Im Bereich von literarischen Texten sind bei Alm und Sproat (2005) die Annotatoren Studenten, die einen Kurs über Märchen besucht haben, bei Marchette et al. (2016) sind die Annotatoren berufliche Historiker. Volkova et al. (2010) jedoch greifen auch für diesen Bereich auf Personen ohne besondere Ausbildung zurück. In der vorliegenden Studie besitzen die Teilnehmer keine für die Dramenanalyse qualifizierende Ausbildung. Es sind allesamt deutsch Muttersprachler und Studenten mit einem ersten Hochschulabschluss. Es war nicht im zeitlichen Rahmen mögliche tatsächliche Experten (beispielsweise Germanisten) zu akquirieren. Die Interpretation der Herausforderungen und Probleme bei der Annotation hängt mit diesem Umstand zusammen und muss beachtet werden. Die Annotation ist nicht das Hauptelement der vorliegenden Studie, weswegen vereinzelte Schwächen in Kauf genommen wurden um einen brauchbaren Gold-Standard-Korpus zu generieren. Zukünftige Studien, die sich mehr auf die Annotation von Dramenrepliken fokussieren, können jedoch versuchen, die angesprochenen Punkte zu verbessern.

#### Fragebogen

Um genauere Daten zu den Problemen und Herausforderungen der Annotation eines derartigen Korpus zu erhalten sollten die Teilnehmer nach vollendeter Annotation noch einen Fragebogen ausfüllen. Die erhobenen Daten tragen zur Kontextualisierung und zur Interpretation des Annotationsverhaltens bei. Außerdem können die gewonenen Informationen zur Verbesserung und Optimierung der Annotation in zukünftigen ähnlichen Projekten beitragen.

Der Fragebogen wurde sehr kurz gehalten. Teilnehmer konnten auf einer ordinalen Skala von 1 – 7 Aussagen zur Schwierigkeit und zur Sicherheit bei der Annotation auf verschiedenen Ebenen zustimmen. Es handelt sich um folgende Aussagen:

Die Annotation der Repliken fiel mir insgesamt schwer.

Die Annotation der Repliken bezüglich der Polarität (Positiv vs Neutral vs Gemischt vs Negativ) fiel mir schwer.

Die Annotation der Repliken bezüglich der Emotionskategorien (Zorn, Traurigkeit etc.) fiel mir schwer.

Ich war mir bezüglich meiner Zuweisungen insgesamt stets sicher.

Ich war mir bezüglich meiner Zuweisungen für die Polaritäten (Positiv vs Neutral vs Gemischt vs Negativ) stets sicher.

Ich war mir bezüglich meiner Zuweisungen für die Emotionskategorien (Zorn, Traurigkeit, etc.) stets sicher.

Ferner gaben die Teilnehmer noch an wie viel Zeit sie, nach eigener Einschätzung, für die Annotation benötigt haben und konnten in einem offenen Antwortfeld die wichtigsten Probleme und Schwierigkeiten eintragen.

Den vollständigen Fragebogen und die Ergebnisse findet man im Anhang. In Kapitel XX werden die Daten des Fragebogens knapp ausgewertet.

### Ergebnisse

Die Ergebnisse unterteilen sich in drei größere Abschnitte: Die Analyse des Übereinstimmungsmetriken der Annotation, also in welchem Ausmaß die Annotatoren bezüglich der erhobenen Daten übereinstimmten und welche Besonderheiten auffallen; die Analyse der Annotationsverteilungen, also wie das Korpus aus Sicht der Annotatoren bezüglich des Sentiments (hier Polarität und Emotionen) konstituiert ist. Als letzter wird der in Kapitel XX kurz besprochene Fragebogen ausgewertet. Alle Auswertungen finden sich als Tabellen, Grafiken und SPSS-Dateien im Anhang. Im Folgenden kann nur der für das weitere Verständnis wichtigste Teil der Ergebnisse erläutert werden.

#### Datenaufbereitung

Die Annotationen der einzelnen Annotatoren wurden manuell aus dem jeweiligen Word-Dokument in eine tabellarische Form (Google Tables) übertragen und verarbeitet. Zur Berechnung einiger Übereinstimmungs-Metriken sowie zur sytematischen Aufbereitung der Daten wurde ein Pyhton-Programm entwickelt, das in Kapitel XX genauer beschrieben wird. Die Ergebnisse wurden ebenfalls mittels Google Tables-Tabellen verwaltet und aufbereitet. Die Ergebnisse für die Sentiment-Verteilung und die Fragebogen-Auswertung wurden über die Statistik-Software IBM Statistics SPSS produziert.

Im folgenden Abschnitt werden einige wesentliche Datenbenennungen und Transformationen zum Verständnis der nachfolgenden Ergebnisaufbereitungen geschildert. Zentrale Variablen sind die Annotationen der Annotatoren pro Replik. Die grundsätzliche Annotation auf der sechsstufigen Skala wird als *Polarität Standard* bezeichnet. Zur detaillierten Analyse wurden weitere Polaritäts-Metriken auf Basis der Angabe von Polarität Standard konstruiert. Polarität Reduziert (vierfach) hebt die Differenzierung von sehr positiv, positiv und sehr negativ, negativ ab und bildet jeweilige Bewertungen auf eine einzelne Gruppe positiv und negativ ab. Polarität Reduziert (dreifach) bildet zusätzlich alle Bewertungen für neutral und gemischt auf eine zusammengefasste Gruppe neutral/gemischt ab, um so eine Differenzierung zwischen eindeutig polaren und uneindeutigen Annotationen zu ermöglichen. Die binäre Annotation nun wird nicht über statistische Konstruktion der Standard-Skala abgeleitet sondern wurden ebenfalls von den Annotatoren angegeben (Polarität dichotom).

Für die Angaben der Emotionskategorien gilt, dass jede Kategorie als vorhanden oder nicht vorhanden ausgezeichnet ist. Ferner würde noch eine Metrik Emotion(en) vorhanden bestimmt, die als keine Emotion vorhanden oder Emotion vorhanden belegt ist. Wenn eine Replik überhaupt keine Emotionsannotation erhält wird die Variable als keine Emotion vorhanden belegt ansonsten als Emotion vorhanden.

Folgende Tabelle fasst die zentralen Variablen zusammen. Für detailliertere Einsichten wird auf den Anhang verwiesen.

Damit die entwickelten Programme (siehe nächstes Kapitel) korrekt verwendet werden können wurden die nominalen Angaben in den einzelnen Tabellen auf numerische Werte abgebildet und codiert, also zum Beispiel für Polarität auf den Wert 1 für positiv und den Wert 2 für negativ. Für jede Variable wurde so eine Annotatoren-übergreifende Tabelle und txt-Datei erstellt, bei der die Bewertungen für jede Replik pro Zeile zusammengefasst sind; also stets 5 Angaben pro Zeile (für jeden Annotator eine) und insgesamt 200 Zeilen (da 200 Repliken). Ferner wurden alle Daten auch sortiert nach Länge der Replik für spätere Längen-spezifische Berechnungen aufbereitet und gesichert. Alle Tabellen sind im Anhang im Ordner Agreement-Daten einsehbar.

#Längengruppen

#Dramengruppen

#### Entwicklung und produzierte Datenstrukturen

Zur Berechnung differenzierter Übereinstimmungs-Metriken und zur holistischen Analyse dieser wurde ein Pyhton-Programm *agreement\_statistics.py* implementiert. Über die Klasse Agreement\_Statistics bietet das Programm zahlreiche Methoden an zum Einlesen von Annotationen in txt- oder tsv-Form, zum Erstellen von Übereinstimmungsmatrizen, zum Berechnen verschiedener Übereinstimmungs-Metriken und präziser Mehrheitsstatistiken. Das Programm ist auf beliebige Annotationsformen anwendbar, insofern man angibt wie viele Kategorien vorliegen und was der Startwert ist. Eine zentrale Transformation, die über diese Programme durchgeführt wird ist die Umwandlung der oben beschriebenen Annotations-Tabellen in Übereinstimmungsmatrizen. Diese geben für jede Replik an wie viele Annotatoren eine Kategorie eines Annotationstyps auswählten; durch folgende Tabelle lässt sich das Konzept von Übereinstimmungsmatrizen erklären:

Derartige Tabellen sind notwendig für die detaillierte Analyse von Übereinstimmungen und die Kalkulation von Übereinstimmungs-Metriken. Als Metriken für die Übereinstimmung können Fleiss‘ Kappa, Krippendorffs Alpha und die prozentuale Übereinstimmung berechnet werden. Zahlreiche Standard-Statistik-Programme bieten Berechnungen für derartige Metriken nicht an (wie z.B. SPSS) weswegen die Entwicklung derartiger Algorithmen notwendig war. Die einzelnen Funktionen greifen dabei auf die Übereinstimmungsmatrizen zur Berechnung zurück. Die Berechnungen für Fleiss‘ Kappa und die prozentualen Übereinstimmungen wurden selbstständig implementiert. Für Krippendorffs Alpha wurde ein externes freies Python-Programm eingebunden (k\_alpha.py), dessen Korrektheit manuell über einfache Beispiele überprüft wurde. Eine weitere essentielle Funktion des Programms ist die Erstellung von Mehrheitsdateien. Das sind Listen die für jede Replik angeben mit welcher Form von Mehrheit Annotationen abgegeben wurde. Je nach Konstitution der Annotationsvariable kann eine Replik eine Annotations-Mehrheit von 2-5 haben. Für Polarität Standard ist beispielsweise denkbar, dass 2 Annotatoren neutral auswählen und die restlichen ihre Annotationen auf andere Optionen verteilen. Eine 5er-Mehrheit liegt vor, wenn alle Annotatoren die gleiche Annotation auswählen. Es gibt für einige Annotationstypen die Möglichkeit, dass keine Mehrheit vorliegt. In diesem Fall wird eine Replik mit dem Wert -1 gesetzt. Derartige Mehrheitsanalysen auch Maße für den Grad der Übereinstimmungen und werden deswegen in Kapitel XX noch explizit betrachtet. Txt-Dateien mit den Angaben aller Mehrheiten findet man auch im Anhang im Ordner Agreement-Daten. Sie werden vom Programm zur Produktion von Übereinstimmungsdaten genutzt und auch später bei der SPSS-Auswertung von Mehrheitsverteilungen. Die genaue Interpretation und Erläuterung aller Daten und Maße wird noch in Kapitel XX bei der Ergebnispräsentation erläutert.

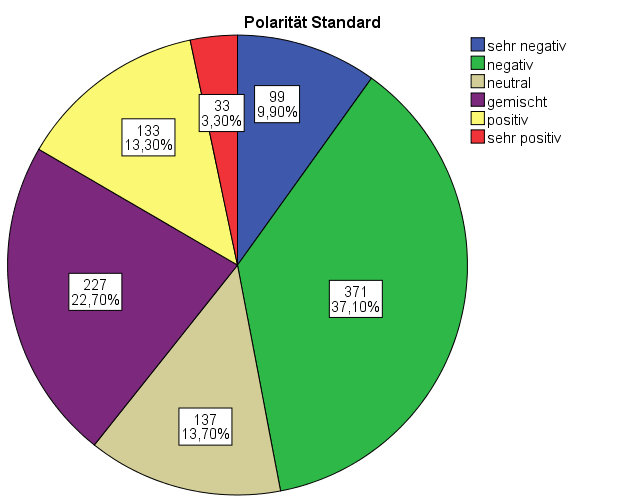
Das Programm ist so aufgebaut, dass es die genannten Metriken für das Gesamt-Test-Korpus aber auch pro Drama und gemäß Längenaufteilungen zur detaillierten Analyse angibt. Die Programme können auch in zukünftigen Annotationsprojekten verwendet werden. Die Berechnungen können per Konsole ausgegeben werden und wurden in Google Table-Dateien übertragen. Diese sind mit alle relevanten Berechnungen und Kalkulationen für alle Maße, gesamt, pro Drama und pro Längenaufteilung im Anhang einsehbar. Für ein tieferes Verständnis der gesamten Funktionalität wird auf den kommentierten Code verwiesen.

#### Sentiment-Verteilungen

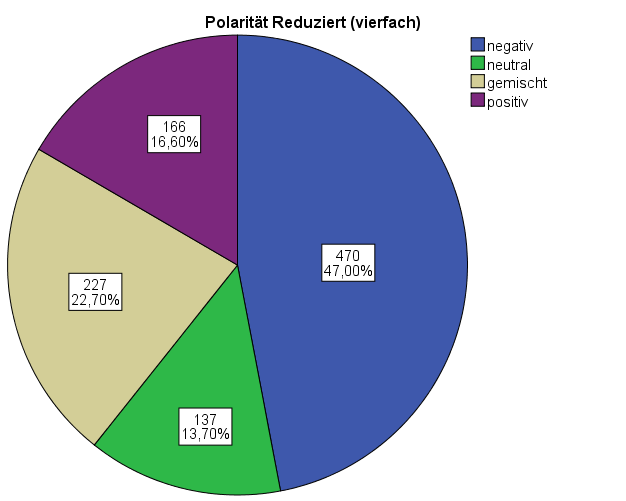
Mittels der von der Annotation manuell aufbereiteten Daten und der Ausgabedaten, die im vorigen Abschnitt beschrieben, wurden, konnten die Sentiment-Verteilungen aus Annotatorensicht deskriptiv untersucht werden. Auf diese Weise kann man das Annotationsverhalten, die Wahrnehmung von Sentiments in Dramenrepliken und die grundsätzliche Sentiment-bezogene Zusammensetzung des Korpus analysieren und Implikationen für die Interpretation der eigenen Daten als auch die Forschung generell formulieren. Die Analyse von Annotationen in der SA ist deswegen ein gängiges Vorgehen, vor allem bei literarischen Texten (Nachweis).

Die zentralen Auswertungen für die Sentiment-Verteilungen wurden mit SPSS durchgeführt und werden nachfolgend präsentiert. Weitere Auswertungen, Grafiken und die Ausgangstabellen pro Annotator und gesamt findet man zur tieferen Analyse im Anhang (z.B. #).

Es wurden von je 200 Repliken von 5 Annotatoren ausgezeichnet, das ergibt insgesamt 1000 Auszeichnungen pro Annotationsmetrik (Polarität Standard, Polarität Dichotom, je Emotionskategorie, Emotion vorhanden). Zunächst werden die Häufigkeitsverteilungen für die Variable Polarität Standard mittels eines Kreisdiagramms betrachtet:

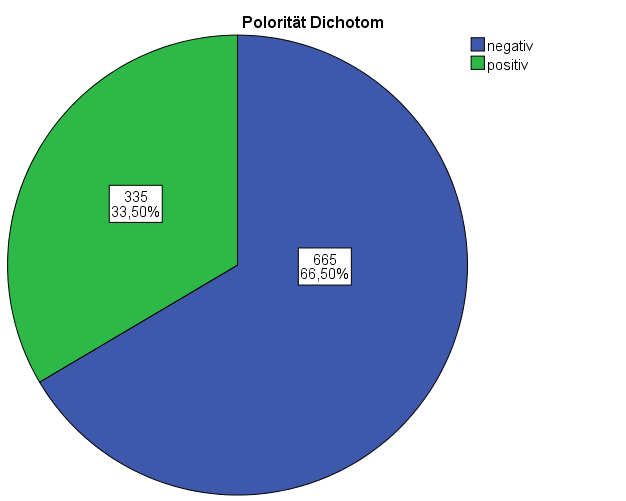


Man erkennt, dass die beiden am häufigsten gewählten Annotationen negativ (37%) und gemischt (23%) sind. Die Extrema der Polaritäten also sehr negativ (10%) und sehr positiv (3%) werden am seltensten ausgewählt. Aus diesem Grund wird als nächstes zur besseren Besprechung der Ergebnisse die Verteilung für Polarität Reduziert (vierfach) betrachtet. Für diese Variable wurden die Gruppen positiv und sehr positiv zu einer Übergruppe positiv zusammengefasst. Selbiges gilt für die Negativitätsgruppen:



Es wird eine besondere Ungleichverteilung deutlich. Die Annotatoren haben fast die Hälfte der Replike und damit die große Mehrzahl als negativ bezüglich ihrer Polarität empfunden. Die nächstgrößere Gruppe ist gemischt (23%) mit. Fast ein Viertel aller Repliken wurden als gemischt, also sowohl positiv als auch negativ wahrgenommen. Verhältnismäßig wenig Repliken wurden tatsächlich als positiv ausgezeichnet (17%). Am seltensten wurden Repliken als neutral bewertet (14%). Dennoch kann man konstatieren, dass sowohl neutral aber vor allem gemischt relevante Klassifizierungsgruppen bei der Annotation von Dramenrepliken sind. Insgesamt wurden etwa 37% der Repliken keiner eindeutigen Polarität (negativ, positiv) zugewiesen. Vor allem die seltenen Annotation mit positiv sind auffällig.

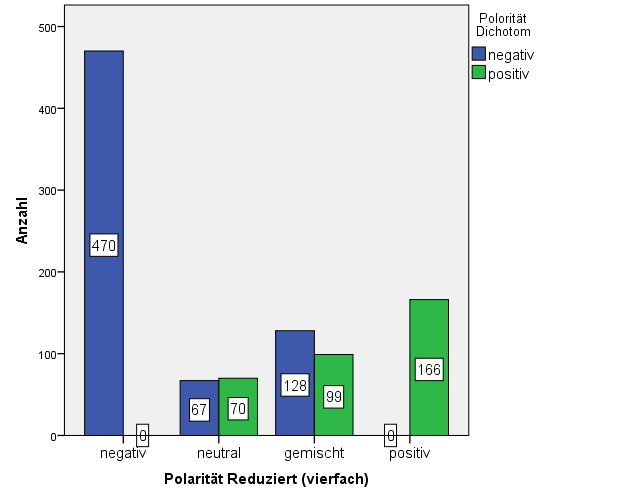
Im Folgenden wird die Verteilung der binären Polaritäts-Annotationen betrachtet. Sollten Teilnehmer eine Replik als neutral oder gemischt eingeordnet haben, wurden sie angewiesen sich für die am ehesten passende Polarität zu entscheiden. Die Verteilung wird durch folgendes Kreisdiagramm illustriert:



Die Ungleichverteilung zwischen negativ und positiv wird bei der binären Bewertung deutlich. Etwa zwei Drittel aller Annotationen für Repliken sind negativ. Ein Chi-Quadrat-Test zur Untersuchung dieser Verteilung zeigt auch statistisch, dass die Polarität signifikant (p<.001) ungleich verteilt ist (siehe Anhang). Diese Ungleichmäßigkeit hat Auswirkungen auf die spätere Berechnung der Übereinstimmungs-Metriken, aber auch auf die Evaluationsberechnungen der SA-Verfahren und der Interpretation dieser.

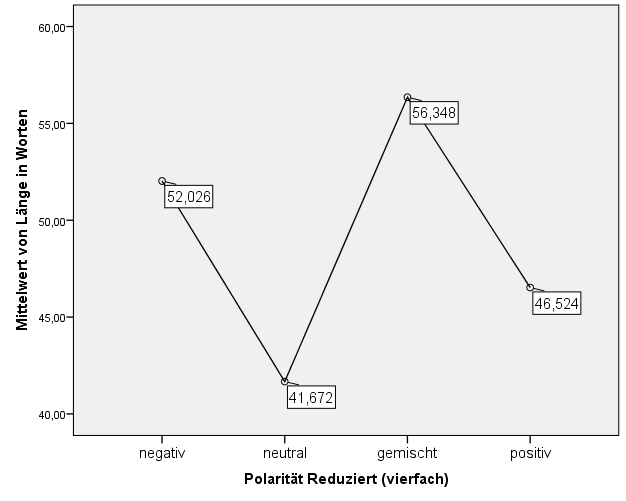
Folgende Kreuztabelle und Balkendiagramm illustriert wie sich die Gruppen neutral und gemischt auf die Polaritätsklassen verteilen:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Kreuztabelle Polarität Reduziert (vierfach)\*Polorität Dichotom** | | | | |
| Anzahl | | | | |
|  | | Polorität Dichotom | | Gesamtsumme |
| negativ | positiv |
| Polarität Reduziert (vierfach) | negativ | 470 | 0 | 470 |
| neutral | 67 | 70 | 137 |
| gemischt | 128 | 99 | 227 |
| positiv | 0 | 166 | 166 |
| Gesamtsumme | | 665 | 335 | 1000 |



Man erkennt, dass die Polaritätsangaben sich ungefähr gleichmäßig auf die neutral/gemischt-Annotationen verteilen, d.h. die Teilnehmer haben vorher als neutral oder gemischt annotierte Repliken bei der binären Entscheidung etwa gleichmäßig als negativ oder positiv annotiert. Bei der Gruppe neutral ist die Verteilung fast exakt gleich, bei gemischten werden mehr Repliken als eher negativ denn als positiv empfunden.

Es wurde ferner noch inferenzstatistisch untersucht ob ein Zusammenhang zwischen der Länge einer Replik und der Annotation besteht. Über Mittelwertvergleiche mit einer einfaktoriellen Varianzanalyse wurde dieser Zusammenhang analysiert und konnte als signifikant festgestellt werden (p<.001). Das folgende Liniendiagramm für die Variable Polarität Reduziert (vierfach) illustriert die durchschnittlichen Längenverteilungen:



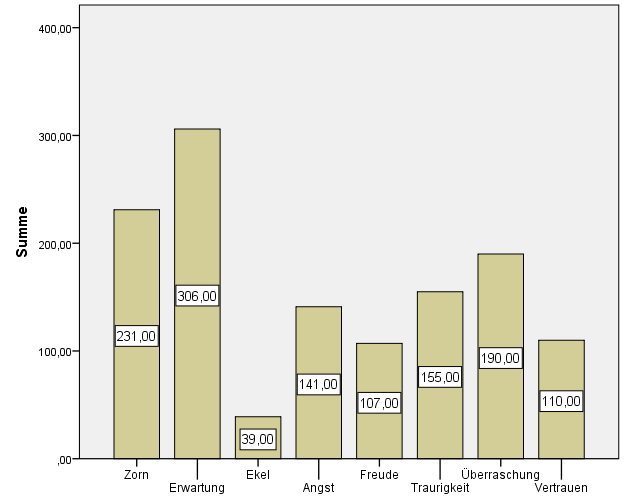
Die Analyse zeigt, dass gemischte Repliken am längsten sind (M=56 Wörter) während neutral bewertete Repliken im Schnitt eher sehr kurz sind (M=42 Wörter). Bei positiven und negativen Repliken fällt auf, dass durchschnittlich längere Repliken als negativ annotiert werden als positive. Insgesamt sind die absoluten Unterschiede aber geringer als beim Vergleich zwischen gemischt und neutral. Auf ähnliche Weise wurde ein Einfluss der Länge auf die dichotome Ausprägung der Polaritätsannotationen untersucht, hierbei konnte jedoch kein signifikantes Ergebnis identifiziert werden. Positiv annotierte Repliken sind im Schnitt genauso lang wie negative (siehe Anhang).

Als nächstes werden die Verteilungen für die Emotionskategorien betrachtet.

|  |  |  |
| --- | --- | --- |
| Emotion | Vorhanden | Nicht Vorhanden |
| Zorn | 231 (23,1%) | 769 (76,9%) |
| Erwartung | 306 (30,6%) | 694 (69,4%) |
| Ekel | 39 (3,9%) | 961 (96,1%) |
| Angst | 141 (14,1%) | 859 (85,9%) |
| Freude | 107 (10,7%) | 893 (89,3%) |
| Traurigkeit | 155 (15,5%) | 845 (84,5%) |
| Überraschung | 190 (19,0%) | 810 (81,0%) |
| Vertrauen | 110 (11,0%) | 890 (89,0%) |

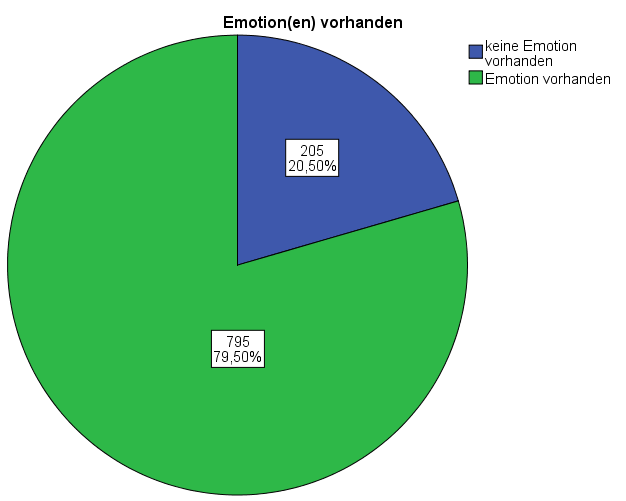
Insgesamt erkennt man für jede Emotion, dass Repliken im Einzelnen deutlich häufiger nicht mit einer Emotion assoziiert werden. Man kann diesbezüglich auch von einer Ungleichverteilung sprechen pro Emotion sprechen, ein Chi-Quadrat-Verteilungstest belegt auch für jede Emotion diese Ungleichverteilung als signifikant (p<.001). Obschon dieser Umstand erwartungskonform ist, ist diese Ungleichverteilung wichtig für die spätere Analyse und Kalkulation der Übereinstimmungsmetriken.

Folgendes Balkendiagramm illustriert die Emotionsannotation im Vergleich:

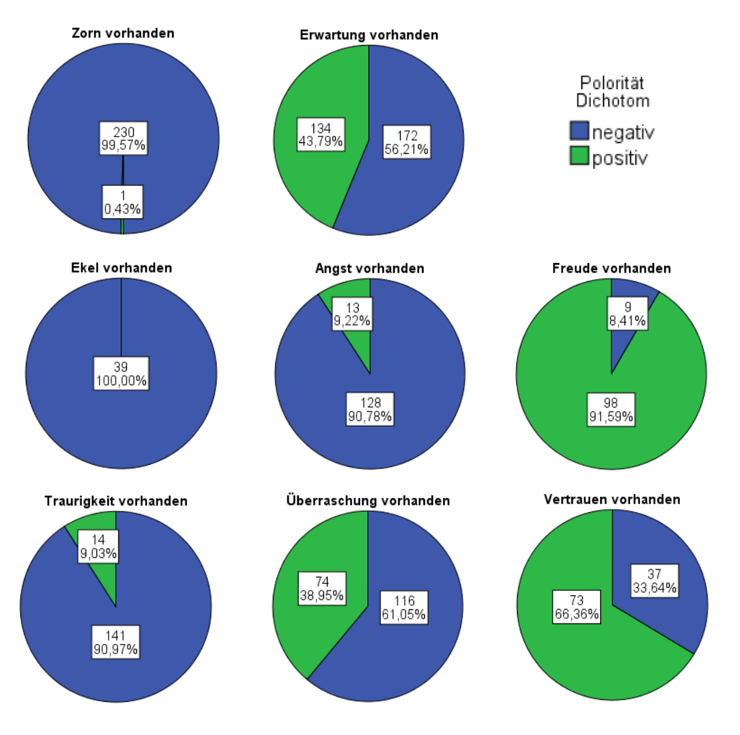


Man erkennt, dass die häufigsten annotierten Emotionen Erwartung und Zorn sind. Eher selten werden Repliken mit der Emotion Freude und Vertrauen annotiert. Nur in 39 von 1000 Fällen wurde eine Replik mit Ekel assoziiert, was somit die seltenste Emotionsannotation darstellt.

Analysiert man jedoch nicht die Assoziation mit einer singulären Emotion sondern ob eine Emotionskategorie überhaupt bezüglich einer Replik aus Sicht der Annotatoren vorliegt, stellt man fest, dass insgesamt deutlich weniger Repliken als vollkommen emotionslos ausgezeichnet werden wie folgendes Kreisdiagramm illustriert:

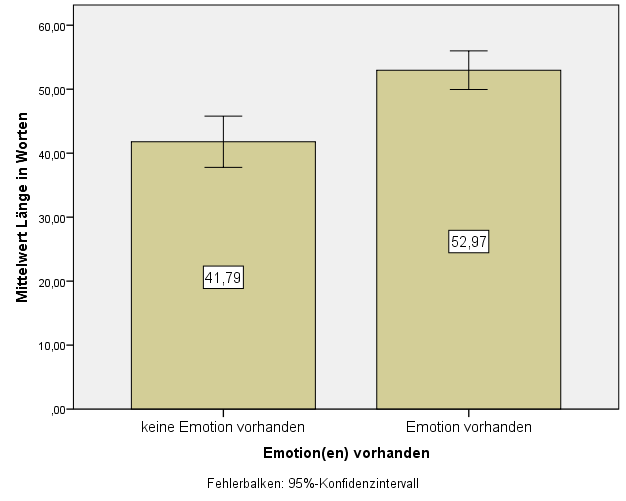


Ferner wurde noch über Kreuztabellen und Häufigkeitsvisualisierungen untersucht wie der Zusammenhang zwischen dem Vorhandensein einzelner Emotionen und der Polarität verläuft (siehe Anhang). Diesbezüglich wird an dieser Stelle eine Sammlung von Kreisdiagrammen präsentiert, die zeigt wie Repliken, die mit einer Emotion annotiert wurden bezüglich Polarität ausgezeichnet wurde. Hier wird die Variable Polarität Dichotom gewählt (positiv, negativ):



Man erkennt für einige Emotionen deutlich Verteilungsunterschiede. Zorn, Ekel, Angst und Traurigkeit werden fast ausschließlich mit Repliken assoziiert, die auch als negativ annotiert wurden. Freude wird tritt fast nur bei positiv ausgezeichneten Repliken auf. Die Emotionskategorien Vertrauen, Überraschung und Erwartung werden als ambivalent bezüglich der Polarität wahrgenommen. Erwartung erscheint fast gleichverteilt.

Es wurde noch, analog zur Polarität, untersucht, ob ein Zusammenhang mit der Länge einer Replik und der Annotation mit einer Emotion besteht. In der Tat können gruppenbasierte Mittelwertvergleiche mit T-Tests zeigen, dass gerade die weitere oben als nicht-ambivalent identifizierten Emotionen einen signifikanten Zusammenhang mit der Länge einer Replik aufweisen. Die negativen Emotionen (Zorn, Ekel, Angst, Traurigkeit) erscheinen eher in längeren Repliken, die positive konnotierte Emotion Freude in kürzeren Repliken. Die Daten können im Anhang im Detail eingesehen werden. Insgesamt kann man aber über einen T-Test mit der Variable Emotion(en) vorhanden und der Länge der Repliken konstatieren, dass Emotionen eher bei längeren Repliken auftreten. Folgendes Balkendiagramm illustriert diesen Umstand:



Es sei noch erwähnt, dass im Gegensatz zur Polaritätsannotation für die Emotionen sehr unterschiedliche Annotationsverhalten pro Annotator festgestellt wurden. Ein Annotator zeichnete fast alle Repliken mit mindestens einer Annotation aus und verwendete dabei häufig die Annotation Erwartung, manche Annotatoren verwendeten nur selten Annotationen. Die Unterschiede kann man sich im Detail im Anhang ansehen.

#### Annotatoren-Übereinstimmung

Im Folgenden wird nun das Ausmaß der Übereinstimmung zwischen den Annotatoren statistisch analysiert. Es handelt sich um einen geläufigen Schritt in der SA bei der Erstellung von Gold-Standards um das Ausmaß der Sicherheit der Annotation zu diskutieren. Man befasst sich auf diese Weise auch mit der grundsätzlichen Frage ob und in welchem Ausmaß man Dramen-Repliken gesichert Sentiments zuweisen kann und wie sehr dies von der individuellen Betrachtung abhängt. Je geringer die Übereinstimmung desto kritischer müssen die finalen Annotationen bei der Nutzung des GS betrachtet werden. Ferner ist die Annotatoren-Übereinstimmung wichtig für die Interpretation der späteren SA-Evaluation. Manche Projekte verwenden beispielsweise Übereinstimmungsmaße als Benchmark für die eigene SA.

#eher nach oben zur Motivation

Alle Übereinstimmungsmaße wurden mittels den in Kapitel XX beschriebenen Datenstrukturen und Programmen berechnet. Google Tables und SPSS wurden für weiterführende Analysen und Illustrationen verwendet.

Es werden verschiedene Übereinstimmungs-Metriken betrachtet. Cohens Kappa ist ein Übereinstimmungsmaß, dass bei der Analyse von zwei Bewertern genutzt werden kann. In der SA-Forschung werden bei mehr als zwei Annotatoren häufig kreuzweise Cohens Kappa-Berechnungen durchgeführt und möglicherweise Durchschnittswerte gebildet (Nachweis). Dies erlaubt zwar einen detaillierten Vergleich aller Annotatoren untereinander, gibt aber kein klares Abbild über die Gesamt-Übereinstimmung aller Annotatoren. Aus diesem Grund wird in der kommenden Auswertung auf Übereinstimmungsmaße für mehr als zwei Annotatoren zurückgegriffen: Fleiss‘ Kappa (Fleiss, 1971) und Krippendorffs‘ Alpha (Nachweis). In der SA-Forschung werden beide eingesetzt. Für alle nachfolgenden Berechnungen wurden auch immer beide Maße kalkuliert und sind im Anhang einsehbar. Da Krippendorffs‘ Alpha (K-Alpha) jedoch als das stabileres Maß der beiden gilt (Nachweis) beschränkt man sich in den nächsten Abschnitten auf eben dieses Maß. In der Tat konnte man feststellen, dass die konkreten Werte beider Maße meist fast exakt gleich sind. Nach Landis und Koch (1977) werden beide Übereinstimmungsmaße nach folgendem Schema interpretiert (hier am Beispiel K-Alpha: α).

α < 0 = schlechte Übereinstimmung

0 < α < 0.2 = schwache Übereinstimmung

0.2 < α < 0.4 = mittelmäßige Übereinstimmung

0.4 < α < 0.6 = moderate Übereinstimmung

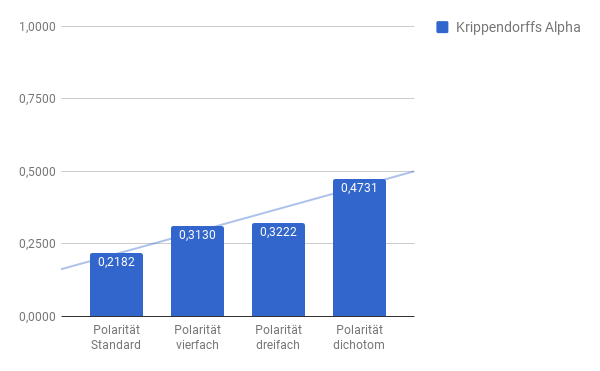
0.6 < α < 0.8 = substantielle Übereinstimmung

0.8 < α < 1 = fast perfekte Übereinstimmung

Beide Maße haben jedoch Probleme wenn die Annotationsausprägungen sehr ungleich verteilt sind. Dies ist beispielsweise bei der Polarität der Fall, da überproportional mehr negative Annotationen verteilt wurden als positive. Ähnliches gilt für die Emotionskategorien, die häufiger nicht vorhanden als vorhanden sind (siehe Kapitel XX). Derartige Ungleichverteilungen führen bei K-Alpha und Fleiss‘ Kappa zu starken Fehlkalkulationen (Nachweis), je nach Ausmaß der Ungleichverteilung und je kleiner die Stichprobe der zu bewertenden Instanzen ist. Um eine korrekte Interpretation zu ermöglichen wird deswegen auch noch stets die prozentuale Durchschnittsübereinstimmung angegeben. Dazu wird für jedes Annotatoren-Paar berechnet zu welchem Anteil diese in der Bewertung exakt übereinstimmen. Die erhaltenen Werte für jedes Annotatoren-Paar werden dann gemittelt. Auch hierbei handelt es sich um eine oft eingesetzte Metrik in der SA um die Übereinstimmung deskriptiv zu beschreiben.

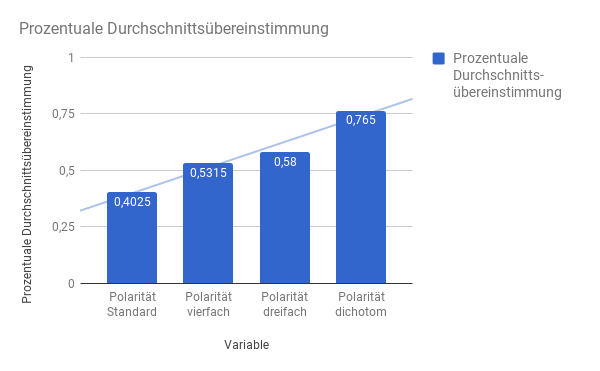
Ferner werden, wenn angebracht, zur weiteren Vertiefung der Daten Mehrheitsstatistiken angegeben. Es können pro Metrik verschiedene Mehrheitstypen vorliegen, generell können für jede Metrik mindestens 5er – 3er-Mehrheiten vorliegen. Bei einer 5e-Mehrheit liegt eine absolute Übereinstimmung bezüglich einer Replik vor, bei einer 3er-Metrik eine schwächere Mehrheit. Bei manchen mehrstufigen Sentiment-Metriken wie Polarität Standard können auch noch schwächere 2er-Mehrheiten vorliegen oder auch keine Mehrheit vorliegen wenn die Replik gleichmäßig bezüglich einer Sentiment-Ausprägung annotiert wurde.

Als erstes wird nun die Übereinstimmung für die Polaritäten in Form von K-Alpha betrachtet. Dazu wird folgendes Balkendiagramm präsentiert, dass die Übereinstimmung je nach gewählter Variable aufzeigt und den diesbezüglichen Trend verdeutlicht:



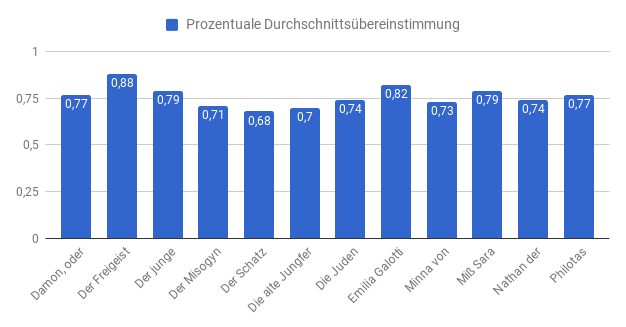
Betrachtet man die differenzierte fünf-wertige Polaritätsannotation stellt man fest, dass die Übereinstimmung nur knapp mittelmäßig ist gemäß K-Alpha (0,22). Die Übereinstimmung steigert sich jedoch stark mit Zusammenführung von Unterkategorien der Polaritäten zu Oberkategorien. Bei der dichotomen Angabe, die lediglich zwischen positiv und negativ unterscheidet ist die Übereinstimmung moderat (0,47). Dennoch muss man festhalten, dass keine substantiellen oder nahezu perfekten Übereinstimmungen für die Polarität vorliegen.

Die durchschnittliche prozentuale Übereinstimmung bestätigt obige Maße auch:



Bei Polarität Standard stimmen die Annotatoren im Schnitt bei etwa 40% der Repliken in der Annotation überein. Hebt man die Differenzierung zwischen sehr positiv und positiv sowie sehr negativ und negativ, steigert sich der Wert auf über 50%. Bei der binären Polarität stimmen Annototaren im Schnitt bei etwa 77% aller Repliken überein.

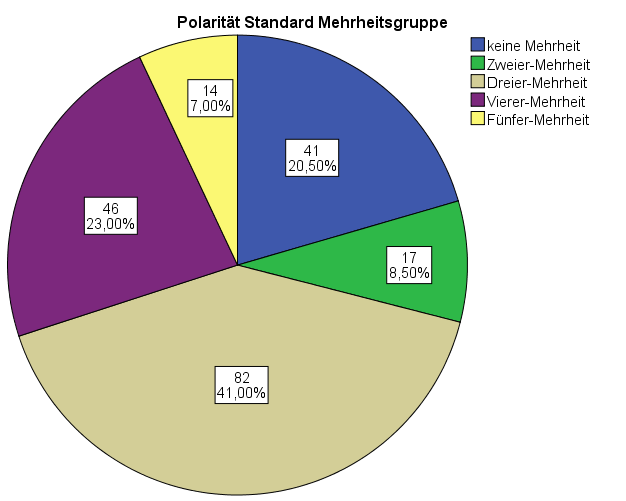
Alle Übereinstimmungsmaße wurden nicht nur bezüglich des Gesamtkorpus berechnet und analysiert sondern auch pro Repliken eines einzelnen Dramas und aufgeteilt in Längen-Gruppen (siehe Kapitel XX). In der Tat können für die Polaritäten jedoch nur vereinzelt besondere dramenspezifische Unterschiede festgestellt werden. Aufgrund der verhältnismäßig geringen Replikenmenge pro Drama ist ferner die Kalkulation der Übereinstimmungsmaße Fleiss‘ Kappa und K-Alpha fehlerbehaftet. An dieser Stelle wird demnach lediglich beispielhaft die prozentuale Durchschnittsübereinstimmung für die zentrale Variable Polarität Dichotom angegeben:



Man erkennt, dass die prozentualen Übereinstimmungen zwischen 68% für das Drama Der Schatz und 88% für das Drama der Freigeist liegt. Es lässt sich jedoch keine besondere Auffälligkeit erkennen, die meisten Dramen haben gleichmäßige Werte zwischen 70 und 80%. Die Dramen sind mit lediglich 6 – 28 Repliken per Drama im Korpus vertreten (je nach Länge des Dramas), weswegen geringfügige Schwierigkeiten bei der Annotation einzelner Repliken die vorliegenden Schwankungen erklären. Insgesamt kann nach der Analyse aller Übereinstimmungsmaße für alle Polaritäten keine wichtiger dramenspezifischer Zusammenhang erkennen. Keines der Dramen weist übermäßig schlechte oder gute Übereinstimmungen auf. Für detailliertere dramenspezifische Analysen wird auf den Anhang verwiesen.

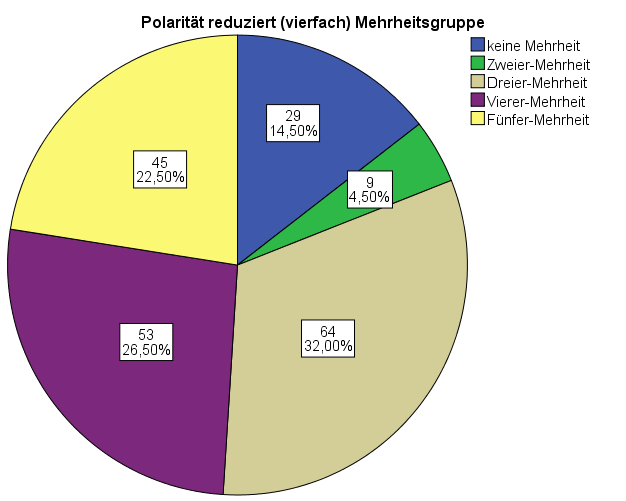
Bezüglich der Längengruppen kann auch kein besonderer Zusammenhang für die Übereinstimmungen festgestellt werden. Unabhängig von Polaritätstyp weisen kurze Repliken keinen besonders anderen Übereinstimmungsgrad auf als längere Repliken, wenn man das Korpus nach Median oder Mittelwert trennt. Die Unterschiede sind marginal und inkonsistent. Für Polarität Standard ist das Ausmaß an Übereinstimmungen bei kürzeren Repliken etwas besser, bei Polarität Dichotom kann man das genaue Gegenteil feststellen. Die Unterschiede sind absolut betrachtet gering, es werden keine Interpretationsgruppen nach Landis und Koch (1977) gewechselt. Auch hier findet man die Daten jedoch auch im Anhang.

Als nächstes werden noch zur Verdeutlichung der Ergebnisse die Mehrheitsgruppen-Verteilung pro Polarität betrachtet und über Kreisdiagramme illustriert:



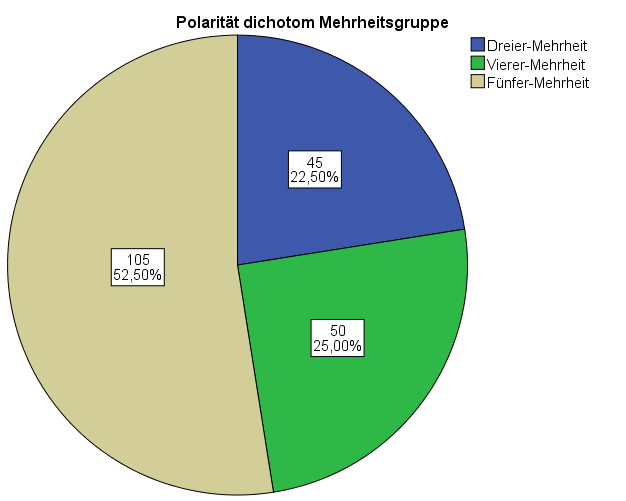
Die häufigste Mehrheitsgruppe für Polarität Standard ist eine 3er-Mehrheit, also Annotatoren waren sich für 41% der Repliken zu dritt bezüglich der Annotation einig. Auffällig ist auch die Häufigkeit der Fälle bei denen keine Mehrheit zustande kam und das sich die Teilnehmer lediglich 5 mal für die Annoation einer Replik einig waren.

Die Verteilung der Mehrheitsgruppen verbessert sich, analog zu obigen Übereinstimmungsmaßen bei der Zusammenfassung von Oberkategorien wie man an Polarität Reduziert (vierfach: positiv, negativ, gemischt, neutral) sehen kann:



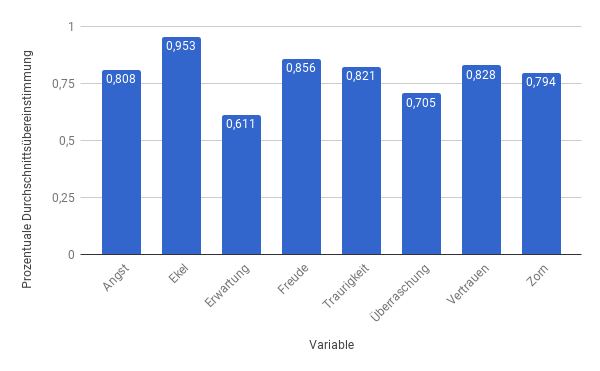
Man kann einen deutlichen Anstieg von absoluten Übereinstimmungen (5er-Mehrheiten) erkennen, aber auch von den anderen höheren Übereinstimmungsgruppen. Dies besagt deutlich, dass sich Personen häufig sicher waren, dass eine Replik negativ bzw. positiv, jedoch uneinig in welchem Ausmaß.

Die zentrale Sentiment-Metrik für diese Studie ist die binäre Polarität. Für diese muss aufgrund der ungeraden Zahl an Annotatoren immer mindestens eine 3er-Mehrheit vorliegen, so dass auf Basis dieser die finale Annotation bestimmt werden kann. Die Verteilung der Mehrheitsgruppen äußert sich dabei wie folgt:



Mehr als die Hälfte der Repliken wurden bezüglich der binären Polarität eindeutig von allen Annotatoren übereinstimmend ausgezeichnet. Mehr als drei Viertel des Korpus wurden mit vollkommenere Übereinstimmung oder zumindest mit einer 4er Mehrheit, bei der sich nur ein Annotator anders entschied für die Polarität, annotiert. Trotzdem muss man anmerken, dass 45 Repliken nur mit einer 3er-Mehrheit bestimmt wurden, die Annotatoren also deutlich unterschiedliche Meinungen bezüglich der Annotation aufwiesen.

Es werden jetzt noch die Übereinstimmungsmaße aller Emotionsannotation betrachtet. Aufgrund der deutlichen Ungleichverteilung dieser, sehr viel häufiger wird eine einzelne Emotion als nicht vorhanden annotiert (siehe Kapitel XX), liefern die trotz hoher prozentualer Übereinstimmungen teilweise schwache Werte. Dies hängt mit bereits angesprochenen Problemen der Maße bei ungleichen Verteilungen zusammen (siehe Kapitel XX). Aus diesem Grund werden hier nur die prozentualen Durchschnittsübereinstimmungen angegeben. Die anderen Maße können jedoch im Anhang eingesehen werden:



Es sind sehr hohe Übereinstimmungswerte erkennbar, d.h. für die große Mehrzahl der Repliken sind sich die Annotatoren einig, dass eine Emotion vorhanden ist oder nicht. Es fällt auf, dass je seltener Emotionen überhaupt annotiert, desto stärker ist die Übereinstimmung. Dies erkennt man an Ekel mit 95% Einigkeit, das gleichzeitig am seltensten insgesamt im Korpus annotiert wurde. Während sich für Erwartung, einer der häufigsten Annotationen, die mit 61% uneiniger waren. Mittelt man alle Übereinstimmungen der Emotionen erhält man einen Wert von 80%.

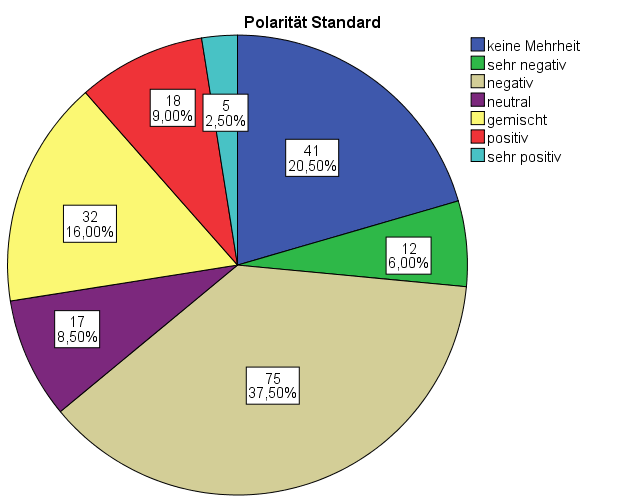
Betrachtet man jedoch jetzt im Detail die Verteilung von Mehrheitstypen pro Kategorie einer Emotionsvariable über Kreuztabellen wird das Gesamtbild der guten Übereinstimmungen relativiert und der eben genannte Zusammenhang zwischen der Häufigkeit einer Emotionsannotation und dem Übereinstimmungsgrad deutlich:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Emotionskategorie/Mehrheitsgruppe | Emotion vorhanden | 3er-Mehrheit | 4er-Mehrheit | 5er-Mehrheit |
| Zorn | nicht vorhanden | 18  11,6% | 28  18,1% | **109**  **70,3%** |
| vorhanden | **20**  **44,4%** | 18  40,0% | 7  15,6% |
| Erwartung | nicht vorhanden | 39  24,7% | 84  53,2% | 35  22,2% |
| vorhanden | 28  66,7% | 10  23,8% | 4  9,5% |
| Ekel | nicht vorhanden | 3  1,5% | 14  7,2% | **178**  **91,3%** |
| vorhanden | 2  40,0% | 2  40,0% | 1  20,0% |
| Angst | nicht vorhanden | 17  9,2% | 51  27,7% | **116**  **63,0%** |
| vorhanden | **9**  **56,3%** | 6  37,5% | 1  6,3% |
| Freude | nicht vorhanden | 17  8,9% | 36  18,9% | **137**  **72,1%** |
| vorhanden | **5**  **50,0%** | 3  30,0% | 2  20,0% |
| Traurigkeit | nicht vorhanden | 21  11,7% | 35  19,6% | **123**  **68,7%** |
| vorhanden | **8**  **38,1%** | **11**  **52,4%** | 2  9,5% |
| Überraschung | nicht vorhanden | 29  15,4% | 90  47,9% | 69  36,7% |
| vorhanden | 8  66,7% | 2  16,7% | 2  16,7% |
| Vertrauen | nicht vorhanden | 18  9,4% | 48  25,0% | **126**  **65,6%** |
| vorhanden | **6**  **75,0%** | 2  25,0% | 0  0,0% |
| Emotion vorhanden | nicht vorhanden | 19  70,4% | 8  29,6% | 0  0,0% |
| vorhanden | 35  20,2% | 46  26,6% | 92  53,2% |

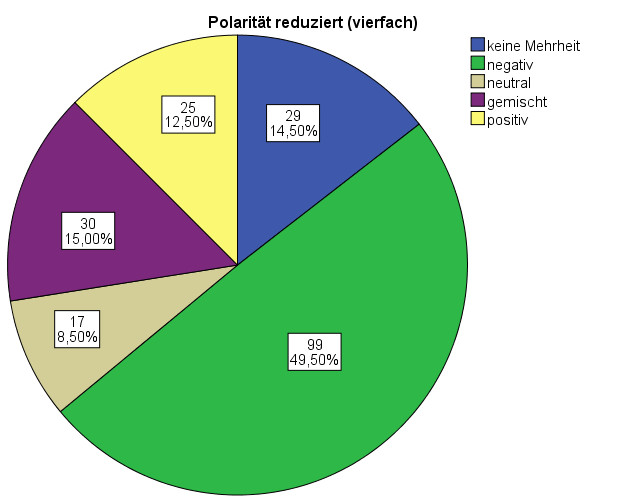
Es werden nicht alle Zusammenhänge im Detail erläutert, jedoch ist bei näherer Analyse für alle Emotionskategorien erkennbar, dass deutlich höhere Mehrheitsklassen vorliegen wenn die Mehrheit der Annotatoren für eine Annotation eine Emotion als nicht vorhanden angibt. Währenddessen sind die Mehrheiten prozentual betrachtet eher schwächere, vor allem Dreier-Mehrheiten, wenn die Emotion als mehrheitlich vorhanden betrachtet wird. Die relevanten Beispiele hierfür wurden fett markiert. Dies besagt insgesamt, dass sich Annotatoren eher übereinstimmen bei Fehlen einer Emotion waren, als wenn diese vorlag. Einzige Ausnahme in diesem Schema bilden die Emotionen Überraschung und Erwartung, bei denen bezüglich der Mehrheitsverteilung, auch bei nicht Vorhandensein einer Emotion gemäß Annotationsmehrheit, kleinere Mehrheitsgruppen wie 3er- und 4er-Mehrheiten vorherrschen. Bezüglich dieser Emotionen sind sich Teilnehmer also insbesondere unsicher.

#### Mehrheitsannotationen

Abschließend werden nun noch die finalen Mehrheitsannotationen beschrieben. Darunter auch die finalen Annotationen für die Variable Polarität Dichotom die in Kapitel XX den Gold Standard und somit die Benchmark für die SA-Evaluation darstellt. Als finale Annotation eine Replik wird diejenige Annotationsausprägung eines Annotationstyps verstanden, die von der Mehrheit der Annotatoren ausgewählt wurde. Dies kann je nach möglichen Ausprägungen bereits bei 2 Annotatoren vorliegen. Bei binären Variablen beispielsweise bei mindestens 3 Annotationen für eine Ausprägung. Bei fünf gleichen Annotationen liegt eine vollständige Übereinstimmung vor. Die so bestimmte Mehrheitsannotation wird als finale Sentiment-Auszeichnung angesehen. Die Verteilung dieser finalen Sentiment-Auszeichnungen werden im Folgenden betrachtet. Sie ähneln grundsätzlich stark den allgemeinen Annotationsverteilungen (siehe Kapitel XX), da die Häufigkeit einer bestimmten Annotationsausprägung trivialerweise damit zusammenhängt, dass sich für diese pro Replik Mehrheiten ergeben. Es werden ferner an dieser Stelle auch noch konkrete Repliken als Beispiele für verschiedene Annotationen gegeben.



Bezüglich Polarität Standard wurde bereits angesprochen, dass sich verhältnismäßig häufig keine Mehrheit für eine Replik gefunden hat (21%). Es bilden sich überwiegend für die Kategorien gemischt und negativ Mehrheiten. Die seltene finale Annotation für Repliken als positiv und sehr positive ist auffällig, es handelt sich um lediglich 23 Repliken die mehrheitlich als positiv betrachtet werden. Dieses Bild bestätigt sich für die vierfache Ausprägung der Polarität also die Zusammenfassung der Positivitäts- und Negativitätsgruppen:



Auch hier fällt das Übergewicht von mit Mehrheiten als negativ annotierten Repliken auf, die fast die Hälfte darstellen. Die zweitgrößte Gruppe sind die Angaben für gemischt und fehlende Mehrheiten.

Zur besseren Veranschaulichung des Korpus seien hier und weiter unten ein paar Repliken als Beispiele für bestimmte Annotationen gezeigt. Zunächst ein Beispiel für eine von allen Annotatoren als neutral identifizierte Replik. Die mittlere fettgedruckte Replik ist die jeweils zu bewertende gewesen:

Der Misogyn 1.Akt, 5.Szene, 11.Replik, Drama-Nummer: 123, ID:56

LELIO:

Gewiß nicht. Aber wieder auf meine Schwester zu kommen – –

WUMSHÄTER:

**Die Ihnen so ähnlich sein soll? Wie ähnlich wird sie Ihnen nun wohl sein? Man wird ohngefähr erkennen können, daß Sie beide aus einer Familie sind.**

LELIO:

Kleinigkeit! Unsere Eltern selbst, konnten uns in der Kindheit nicht unterscheiden, wenn wir aus Mutwillen die Kleider vertauscht hatten.

Hier die einzige Replik, die von allen Annotatoren als „gemischt“ ausgezeichnet wurde:

Nathan der Weise 1.Akt, 4.Szene, 10.Replik, Drama-Nummer: 179, ID:177

DAJA:

Was quält Ihr mich? – Ihr gierig Aug' erriet ihn hinter Den dicht verschränkten Palmen schon; und folgt Ihm unverrückt. Sie läßt Euch bitten, – Euch Beschwören, – ungesäumt ihn anzugehn. O eilt! Sie wird Euch aus dem Fenster winken, Ob er hinauf geht oder weiter ab Sich schlägt. O eilt!

NATHAN:

**So wie ich vom Kamele Gestiegen? – Schickt sich das? – Geh, eile du Ihm zu; und meld' ihm meine Wiederkunft. Gib Acht, der Biedermann hat nur mein Haus In meinem Absein nicht betreten wollen; Und kömmt nicht ungern, wenn der Vater selbst Ihn laden läßt. Geh, sag', ich laß' ihn bitten, Ihn herzlich bitten ...**

DAJA:

All umsonst! Er kömmt Euch nicht. – Denn kurz; er kömmt zu keinem Juden.

Es folgt ein Beispiel einer übereinstimmen als negativ bewerteten Replik:

Der junge Gelehrte 3.Akt, 17.Szene, 3.Replik, Drama-Nummer: 1000, ID:35

CHRYSANDER:

Aber was ist dir denn in den Kopf gekommen?

DAMIS:

**Ich bin es längst überdrüssig gewesen, länger in Deutschland zu bleiben; in diesem nordischen Sitze der Grobheit und Dummheit; wo es alle Elemente verwehren, klug zu sein; wo kaum alle hundert Jahr ein Geist meines gleichen geboren wird – –**

CHRYSANDER:

Hast du vergessen, daß Deutschland dein Vaterland ist?

Abschließend eine Beispiel für eine gemäß Teilnehmern eindeutig positiv konnotierte Replik:

Der Schatz 1.Akt, 5.Szene, 14.Replik, Drama-Nummer: 253, ID:73

MASKARILL:

Ich leihe Ihnen, mein Herr, –

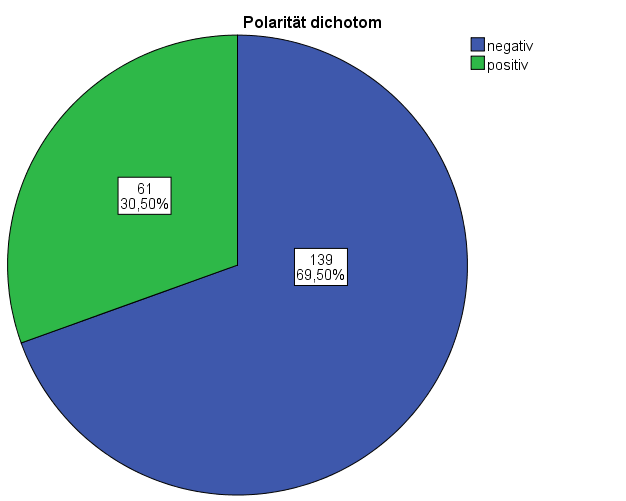
LELIO:

**Sage nicht: mein Herr. Nenne mich deinen Freund. Ich wenigstens will dich Zeit Lebens für meinen einzigen, besten Freund halten.**

MASKARILL:

Behüte der Himmel! Sollte ich, einer so kleinen nichtswürdigen Gefälligkeit wegen, den Respekt bei Seite setzen, den ich Ihnen schuldig bin?

Die finale und wichtigste Annotationsgruppe stellt Polarität dichotom dar:



Die finalen Annotationen für die Polarität Dichotom, basierend auf den Mehrheitsangaben, ergeben 61 positiv annotierte Repliken und 139 negative Repliken. In Kapitel XX wurde bereits der Übereinstimmungsgrad bezüglich dieser Repliken besprochen, der insgesamt moderat ist (K-Alpha=#TODO). Die Ungleichverteilung muss bei der SA-Evaluation sowohl bei der Durchführung als auch bei der Interpretation beachtet werden. Die finalen Annotationen befinden sich als txt-Datei in Form einer Liste im Ordner Agreement-Daten im Anhang (analog zu allen Mehrheitsannotationen).

Zum besseren Verständnis der Annotationsprobleme sei hier noch ein Beispiel für eine Replik genannt die nur mit schwacher Mehrheit (3er-Mehrheit) als negativ annotiert wurde, bei der sich die Annotatoren also deutlich uneinige waren:

Der junge Gelehrte 1.Akt, 6.Szene, 70.Replik, Drama-Nummer: 212, ID:54

ANTON:

Vielleicht; vielleicht nicht. Wenn ich wüßte was er für ein Buch zuletzt gelesen hätte, und wenn ich dieses Buch selbst lesen könnte, und wenn – –

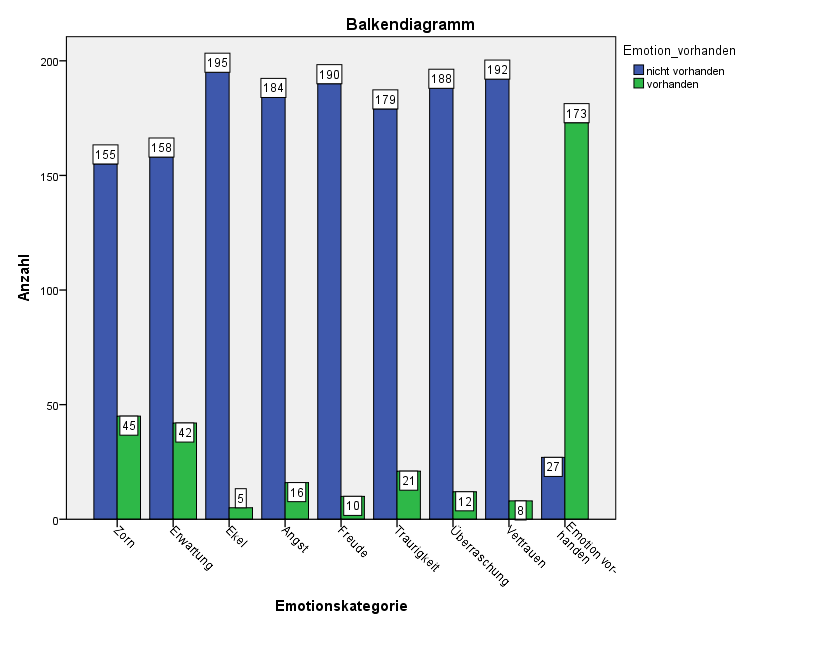
CHRYSANDER:

**Ich sehe schon, ich werde deine Hülfe nötig haben. Du bist zwar ein Gauner, aber ich weiß auch, man kömmt jetzt mit Betriegern weiter, als mit ehrlichen Leuten.**

ANTON:

Ei, Herr Chrysander, für was halten Sie mich?

Anbei sind nun noch die Mehrheitsannotationen der Emotionskategorien über Kreisdiagramme illustriert. Da diese nicht weiter in der SA-Evaluation aufgegriffen werden, werden sie an dieser Stelle nicht näher ausgeführt. Zukünftige Studien können die finalen Annotationen jedoch zur Emotionsbasierten SA-Evaluation nutzen.

Man erkennt über die gruppierten Balkendiagramme, dass sich die Ergebnisse der allgemeinen Häufigkeitsverteilungen verdeutlichen. Obschon in der Mehrzahl der Fälle eine Mehrheit für eine vorhandene Emotion zu finden ist (173), können die einzelnen Emotionen verhältnismäßig selten mehrheitlich einer Replik zugeordnet werden. Am häufigsten gilt dies für die Emotionen Zorn und Erwartung. Für alle anderen Emotionen lassen sich je zwischen 5 (Ekel) – 21 (Traurigkeit) Annotationsmehrheiten konstatieren. Dies verdeutlicht nochmal die geringe Übereinstimmung bei der Annoatation des Vorhandenseins einer Emotion.

#### Fragebogen-Auswertung

Im letzten Ergebnis-Abschnitt bezüglich der Test-Korpus-Annoatation wird nun noch der Fragebogen (siehe Kapitel XX) ausgewertet. Auf diese Weise wurden über verschiedene Fragen Informationen zur Schwierigkeit, Problemen und Herausforderungen bei der Annotation gesammelt.

Zunächst wird die Auswertungsstatistik für die Fragen zur Einschätzung der Schwierigkeit und Sicherheit bei der Annotation präsentiert. Es handelt sich um Zustimmungsfragen auf einer siebenstufigen Skala. Es werden Mittelwert und Median angegeben.

|  |  |  |
| --- | --- | --- |
| **Statistiken** | | |
|  | Mittelwert | Median |
| Die Annotation der Repliken fiel mir insgesamt schwer. | 5,4000 | 6,0000 |
| Die Annotation der Repliken bezüglich der Polarität (Positiv vs Neutral vs Gemischt vs Negativ) fiel mir schwer. | 4,6000 | 5,0000 |
| Die Annotation der Repliken bezüglich der Emotionskategorien (Zorn, Traurigkeit etc.) fiel mir schwer. | 4,6000 | 5,0000 |
| Ich war mir bezüglich meiner Zuweisungen insgesamt stets sicher. | 3,4000 | 3,0000 |
| Ich war mir bezüglich meiner Zuweisungen für die Polaritäten (Positiv vs Neutral vs Gemischt vs Negativ) stets sicher. | 4,0000 | 4,0000 |
| Ich war mir bezüglich meiner Zuweisungen für die Emotionskategorien (Zorn, Traurigkeit, etc.) stets sicher. | 3,4000 | 3,0000 |

Bezüglich der Schwierigkeit stellt man fest, dass Annotatoren mit einem Median von 6 die Annotation als schwer und anspruchsvoll empfanden. Die Statistik zeigt dabei keinen besonderen Unterschied zwischen der Annotation von Polaritäten oder Emotionen. Bezüglich der Sicherheit mit der Annotation kann man konstatieren, dass die Annotatoren sich durchschnittlich sicher waren und etwas unsicherer bezüglich der Polarität als bezüglich der Emotionen. Die Statistik zu den subjektiven Selbsteinschätzung zeigt somit zwar, dass sich die Teilnehmer nicht vollständig unsicher bei der Auszeichnung waren, man aber auch in keiner Weise von hoher Sicherheit und Überzeugung bei der Annotation sprechen kann.

Die zeitliche Beanspruchung aus Sicht der Annotatoren wird durch folgende Tabelle in Minuten illustriert:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Deskriptive Statistiken** | | | | | | | |
|  | N | Bereich | Minimum | Maximum | Mittelwert | Standardabweichung | Varianz |
| Zeit in Minuten | 5 | 210,00 | 180,00 | 390,00 | 306,0000 | 90,99451 | 8280,000 |

Im Schnitt benötigten Teilnehmer, laut eigener Einschätzung, 306 Minuten, also 5,1 Stunden zur Annotation von 200 Repliken. Am schnellsten arbeitet ein Annotator mit 3 Stunden, am längsten benötigte ein Annotator mit 6,5 Stunden.

Ferner gaben die Teilnehmer noch in einer offenen Frage die wichtigsten Problem bei der Annotation an. Die genauen Aussagen können dem Anhang entnommen werden. Die wichtigsten Punkte werden hier stichpunktartig zusammengefasst:

* schwer verständlicher Sprachstil
* veraltete Sprache
* Schwierigkeit eine Replik in einen Gesamtkontext einzuordnen
* Umgang und Verständnis von Ironie
* Emotionen und Polaritäten ändern sich während einer Replik häufig, vor allem bei längeren Repliken
* Manche Repliken erscheinen inhaltsleer, da sie lediglich aus „belanglosen Phrasen“ bestehen, was die Polaritätsbestimmung schwer macht

## Diskussion

# SA-Evaluation

In diesem Kapitel wird die systematische Evaluation aller eingesetzter SA-Methoden erläutert. Zunächst wird das grundsätzliche Vorgehen sowie die zentralen Metriken erörtert. Daraufhin knapp das Programm zur Durchführung der SA-Evaluation beschrieben und dann die Ergebnisse knapp zusammengefasst und ausschnittsweise beschrieben. Die kompletten Ergebnistabellen findet man im Anhang. Abschließend werden die Ergebnisse im Kontext der Forschung kurz diskutiert.

## Vorgehen

Das Evaluationsvorgehen orientiert sich an herkömmlichen Vorgehen in der SA-Forschung (Nachweis). Die SA-Methoden werden systematisch gegen das annotierte Korpus auf ihre Leistung verglichen. Die Sentiment-Metrik für die Evaluation ist die herkömmliche dichotome Polarität (positiv vs negativ). Die finale Gold-Standard-Annotation wird in Kapitel XX beschrieben und basiert auf den Mehrheitsangaben der Annotatoren. Es findet keine Evaluation gegenüber komplexeren Polaritätsmetriken oder auch den Emotionsannotationen statt. Die Übereinstimmungen für die komplexeren Polaritäten waren zu schwach um eine brauchbare Evaluation durchzuführen. Bezüglich der Emotionen liegen diese für eine aussagekräftige Evaluation zu selten mehrstimmig vor (siehe Kapitel XX). Insgesamt ist die vorliegende Studie damit konform zur momentanen Forschung, die auch auf binäre Polaritäten fokussiert ist. Ferner ist dies die erste dem Autor bekannte Studie die für Dramentexte eine systematische Evaluation mit Gold-Standard durchgeführt. Dennoch können zukünftige Studien Ideen und Ergebnisse aufgreifen und mit größeren annotierten Datenmengen auch die genannten komplexeren Kategorien analysieren.

Die Evaluation wurde für alle kombinatorischen Herangehensweisen, der vorgestellten SA-Verfahren, Lexika und Metriken durchgeführt. Diese wurden bereits in Kapitel XX aufgeführt und erläutert. Es handelt sich insgesamt um fünf verwendete Lexika sowie ein kombiniertes Lexikon. Für jedes Lexikon mit gewichteten Polaritätsangaben wurde eine dichotome Term-Zähl-Version gebildet. Für das kombinierte Lexikon wurden zwei Metriken entwickelt. Insgesamt ergibt das 10 Polaritäts-Metriken. Es werden die nicht-normalisierten rohen Polaritäten betrachtet (siehe Kapitel XX). Die Betrachtungsebene ist gemäß GS die Replik. Es kann eine Lexikonerweiterung mittels dem Tool von Jurish (2010) vorliegen (dtaExtended) oder nicht (noExtension). Wahlweise findet keine Lemmatisierung (noLemma), Lemmatisierung nur auf Text (textLemma) oder Lemmatisierung sowohl auf Text als auf Lexikonseite (bothLemma) statt. Es gibt zwei mögliche Lemmatisierer (textblob, treetagger). Es wird auch der Einsatz von drei Stoppwortlisten (standardList, extendedFilteredList, extendedList) bzw. der Nicht-Verwendung von Stoppwortlisten untersucht (noStopwordList). Des Weiteren wurde für den finalen Wortabgleich-Schritt bei der Sentiment-Berechnung zwischen Beachten der Groß- und Kleinschreibung (caseSensitive) und Ignorieren dieser unterschieden (caseInSensitive). Dies ergibt insgesamt 80 unterschiedliche Herangehensweisen pro Polaritäts-Metrik. Insgesamt ergibt dies also 800 kombinatorische SA-Möglichkeiten. Dabei muss man jedoch beachten, dass zwar jede Option unterschiedlich im Detail ist, sich eine Kombinationen (z.B. caseInSensitive vs caseSensitive) sich sehr ähneln. Dennoch kann im Folgenden aufgrund der hohen Kombinatorik nicht die Auswertung jeder Metrik im Detail betrachtet werden. Aus diesem Grund wird nach Lexikon getrennt und sich für jede Metrik ein paar wichtige Beispiele betrachtet und grundsätzliche Erkenntnisse erläutert. Dazu wird noch ein Gesamtfazit formuliert und die besten Herangehensweisen diskutiert. Für eine detaillierte Ergebnisanalyse muss jedoch auf den Anhang verwiesen werden.

Zur Analyse der Evaluation werden zunächst einige wichtige Evaluations-Metriken definiert. Die Metriken werden auch in anderen Studien eingesetzt und erläutert (Nachweis). Zur Erklärungen dieser Maße wird sich an Goncalves et al. (2013) orientiert. Folgende Tabelle definiert einige Grundmaße zur Berechnung der letztendlichen Evaluations-Metriken:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Vorhergesagte Erwartungen (SA-Verfahren) | |  |
| Negativ | Positiv | Gold Standard |
| Tatsächliche Beobachtungen (Gold Standard) | Negativ | A (True Negatives) | B (False Positives) | Alle negativen Repliken (139) |
| Positiv | C (False Negatives) | D (True Positives) | Alle positiven Repliken (41) |
|  | Summe | Alle negativen Vorhersagen | Alle positiven Vorhersagen |  |

Unter A werden alle Repliken verstanden die negativ annotiert sind und als solche auch korrekt als negativ erkannt werden. D bezeichnet das gleiche für den Fall einer positiven Annotation und Erkennung. Über den Fall B werden jene Repliken gezählt die im Gold-Standard negativ sind und fälschlicherweise als positiv annotiert werden. Über die Zelle B die Repliken die tatsächlich positiv sind aber vom SA-Verfahren eine negative Bewertung erhalten. A + C ergibt alle negativen Vorhersagen des SA-Systems, B + D alle positiven. Ferner erhält man über A+B die Zahl aller negativ annotierten Repliken des GS (139) und über C+D alle positiven Repliken.

Das zentrale Evaluationsmaß in der SA ist die Genauigkeit, im Englischen accuracy genannt. Diese ergibt sich über die Formel (A+D)/(A+B+C+D) oder vereinfacht gesagt, der Anteil aller insgesamt korrekt erkannten Repliken an allen Repliken des GS. Meist ist dieses Maß ausreichend um die Leistung zu messen und wird auch in der Forschung für Studienübergreifende Vergleiche herangezogen. Die Genauigkeit beträgt 1 wenn alle Repliken korrekt erkannt werden und 0 wenn keine korrekt erkannt wird.

Wenn man jedoch detaillierte Analysen bezüglich Unterschieden und Besonderheiten in der Erkennungsleistung für speziell positive und speziell negative GS-Einheiten wird auf die Maße Recall, Precision und den F-Wert zurückgegriffen. Alle drei werden separat für positive und negative Repliken berechnet. Der Recall für negative Repliken ist definiert als A/(A+B) also der Anteil korrekt erkannter negativer Repliken an allen negativen Repliken im GS. Ähnlich ist der Recall für positive Repliken als D/(D+C) definiert. Es ist also ein der Genauigkeit ähnliches Maß, jedoch speziell nur für positive und negative Einheiten. Der Wertebereich verläuft wieder von 0 – 1. Mit einem Wert von 1 werden alle Einheiten der Kategorie korrekt erkannt. Precision für die negativen Vorhersagen wird kalkuliert als A/(A+C) und für positive Vorhersagen als D/(D+B). Das Maß gibt also den Anteil korrekt erkannter Repliken an allen für eine Kategorie ausgezeichneten Repliken an. Über Precision können somit problematische SA-Verfahren erkannt werden die eine Polarität übermäßig ausgeben. Auch diese Metrik verläuft von 0 – 1. Der F-Wert kombiniert Precision und Recall um eine erleichterte Gesamtinterpretation zu ermöglichen: 2 x (Precision \* Recall)/(Precision + Recall). Der F-Wert verläuft ebenfalls von 0 – 1 mit einer perfekten Erkennungsleistung für eine Kategorie bei einem Wert von 1 für Recall und Precision. Die Erweiterung des zentralen Evaluationsmaß Genauigkeit mit Recall, Precision und F-Wert ist in der vorliegenden Studie zwingen notwendig da die Polaritäten sehr ungleich verteilt sind. Ein SA-Verfahren, dass beispielsweise alle Repliken als negativ bewertet würde im vorliegenden GS eine gute Erkennungsleistung von ca. 0,7 produzieren (da etwa 70% der Repliken negativ sind). Die Fehlerhaftigkeit dieses Verfahren lässt sich nur durch kategorienspezifische Maße wie die genannten Recall, Precision und F-Wert erkennen. Außerdem können die Verfahren somit detailliert und erschöpfen ausgewertet werden.

Folgende Tabelle fasst nochmal alle Maße zusammen:

#TODO Tabelle

## Entwicklung

Zur Durchführung der SA-Evaluation für alle SA-Verfahren und zur systematischen Berechnung aller Evaluationsmaße wurde ein Python-Programm entwickelt: evaluation\_test\_corpus\_analysis.py. Das Programm bietet verschiedene Klassen und Methoden zum Abgleich aller SA-Verfahren gegenüber einem Gold-Standard und zur Ausgabe aller Evaluationsinformationen in verschiedenen Formaten. Dieses Programm stellt ein allgemeines SA-Evaluationsframework dar und kann mit verschiedenen Gold-Standards-Korpora und SA-Verfahren genutzt und erweitert werden. Damit kann es in zukünftigen Projekt zur Evaluation eingesetzt werden.

Zentral für die Ausführung aller Operationen ist die Klasse Test\_Corpus\_Evaluation. Die Klasse besitzt Methoden zur automatischen Kalkulation und Ausgabe von Evaluations-Metriken und -Informationen für einzelne SA-Methoden aber auch alle SA-Methoden gleichzeitig. Eine zentrale Datei, die dabei genutzt wird befindet sich im Ordner Evaluation/Test-Korpus-Evaluation/Benchmark-Daten/Polaritaet\_dichotom.txt. Es handelt sich dabei um den Gold-Standard, der hier zeilenweise notiert und eingelesen wird mit den numerischen Werten 1 für negativ und 2 für positiv. Die Klasse Test\_Corpus\_Evaluation greift auf Objekte der Klasse Comparison\_Result\_Polarity zurück. Diese Klasse stellt Datenstrukturen zur Speicherung aller Evaluations-Metriken sowie Methoden zur Berechnung dieser, aber auch zur Berechnung von Kreuztabellen nach obigem Schema (siehe Kapitel XX). Innerhalb von Test\_Corpus\_Evaluation wird der Gold-Standard als Liste initialisiert sowie der als Pickle-Datei abgespeicherte Test-Korpus (wie in Kapitel XX) geladen. Über die zentrale Methode attachSentimentInfoOnTestCorpus werden Objekte der Klasse Sentiment\_Analyzer gemäß ausgewählter SA-Verfahrens-Kombinatorik erstellt, die SA durchgeführt und das Polaritätsergebnis an den Test-Korpus intern angefügt. Anschließend können über die Methode comparePolarityMetricWithBenchmark die Ergebnisse als Comparison\_Result\_Polarity-Objekte erstellt werden. Als zentrale Ausführungsmethoden seien noch setEvaluationInfoOfAllCombinationsForSingleMetric genannt, dass für eine angegebene Metrik für alle 80 Herangehensweisen den kompletten Ergebnisoutput produziert sowie createAllOutputsOfAllMetrics, dass für alle 10 Metriken den gesamten Output für alle Herangehensweisen erstellt. Die gesamten Berechnungen und Abfolgen werden hier sehr vereinfacht dargestellt. Für weitere Informationen sei auf den kommentierten Programm-Code verwiesen.

Der Output wird im Ordner Evaluation-Results abgespeichert. Es können detaillierte txt-Dateien für jede einzelne Herangehensweise produziert werden sowie tabellarische Ergebniszusammenfassungen für alle Maße im tsv-Format. Die detaillierten Dateien enthalten den zusammengesetzten Namen der SA-Methode, die Hauptmaße (Genauigkeit, Recall Negativ usw.) sowie eine Kreuztabelle mit den genauen Angaben für True Positives, False Positives, True Negatives und False Negatives. Danach folgen die Detailergebnisse für alle Repliken einzeln aufgelistet, also was die SA-Methode im Vergleich zur Benchmark angibt und welche Wörter mit welchen Polaritätsangaben exakt als SBWs erkannt wurden. Über diese Angaben konnten exakte Analysen der einzelnen SA-Methoden durchgeführt werden.

#Beispiele?

Die produzierten tsv-Dateien geben alle Hauptmetriken pro Polaritätsmetrik an. Die Benennungen der tsv-Dateien setzen sich zusammen aus dem Namen der Polaritätsmetrik und \_majorMetrics.tsv. Sie sind so formatiert, dass per Zeile der Kombinationstyp als tab-getrennte Liste mit den Methodenausprägungen und den Methodennamen am Anfang angegeben wird, also zum Beispiel:

noExtension\_treetagger\_bothLemma\_standardList\_caseInSensitive \t noExtension \t treetagger \t bothLemma \t standardList \t caseInSensitive

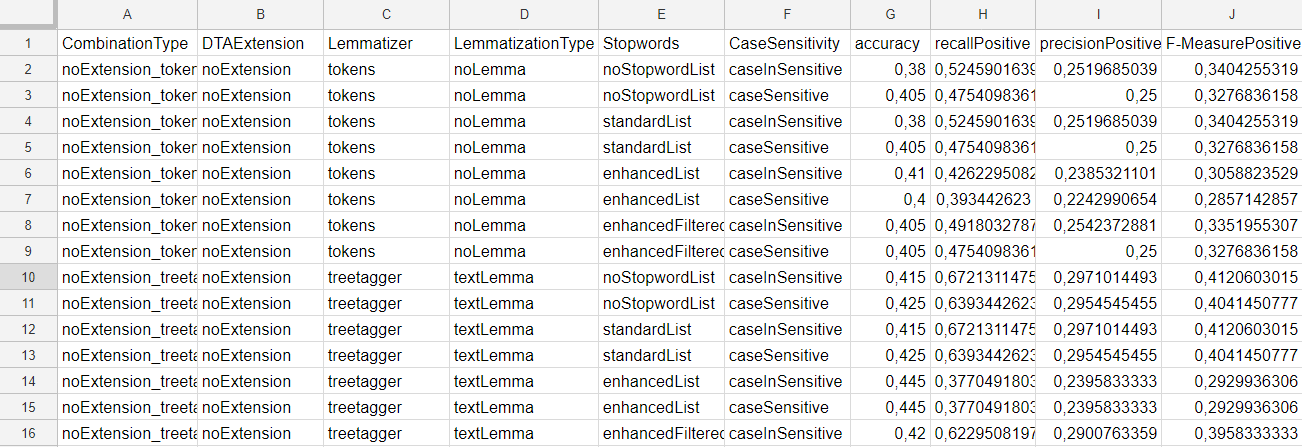
Darauf folgen in der gleichen Zeile auch tab-getrennt die Hauptmaße Genauigkeit, Recall Positive, Precision Positive usw. Alle Ausgaben können im Anhang im Ordner Evaluation-Results eingesehen werden. Die tsv-Tabellen wurden, wie im nächsten Kapitel noch beschrieben wird, auch in Google Tables für die weitere Analyse übertragen, auch diese findet man im Anhang.

## Ergebnisse

In diesem Kapitel werden nun die finalen Evaluationsergebnisse zusammengefasst und erläutert. Aufgrund der Menge der Daten werden pro Lexikon und Polaritäts-Metrik nur die wichtigsten Informationen anhand der besten und schlechtesten Evaluationsleistungen angegeben und besprochen. In Kapitel XX werden dann die Ergebnisse auch vergleichend zusammengefasst, zentrale Aussagen formuliert und die besten SA-Verfahren vorgestellt.

### Datenaufbereitung

Die tsv-Tabellen, die über das in Kapitel XX beschriebenen Programm erzeugt wurden, wurden einfach in Tabellen mittels Google Tables übertragen. So wurden pro Polarität alle 80 Herangehensweisen systematisch gesichert und analysiert. Folgender Tabellenausschnitt illustriert grob das Format der Tabellen:



In der ersten Spalte befindet sich der Name des Kombinationstyps, danach folgen die einzelnen SA-Optionen mit ihren jeweiligen Ausprägungen und den konkreten Ergebnissen:

accuracy, recallPositive, precisionPositive, F-MeasurePositive, recallNegative, precisionNegative, F-MeasureNegative, recallAverage, precisionAverage, F-MeasureAverage, truePositives, falsePositives, trueNegatives falseNegatives

Durch diese spaltenweise Ordnung konnten durch Tabellensortierungen die Ergebnisse bezüglich expliziter SA-Optionen untersucht werden. Neben diesen ungeordneten Tabellen wurde noch je eine Tabelle pro Polaritäts-Metrik erstellt die nach accuracy (Genauigkeit) geordnet ist, also mit der besten Leistung oben. Da accuracy die zentrale SA-Metrik ist wurden diese Tabellen für die Ergebnisinterpretation und -analyse genutzt. Dabei wurde auch auf die Besonderheiten der anderen Maße (Recall, Precision. F-Wert) geachtet, die Vorsortierung über accuracy ist jedoch eine erste passende Orientierung zur Identifikation der besten SA-Verfahren pro Polaritäts-Metrik. Ferner wurde eine Zusammenfassungstabelle aller 800 Herangehensweisen erstellt, sowohl ungeordnet als auch sortiert. Zum besseren Vergleich der einzelnen Polaritäts-Metriken wurde ferner auch eine Tabelle aufbereitet die nur aus den je fünf besten Verfahren gemäß accuracy besteht, welche ebenfalls sortiert und unsortiert vorliegt.

Bei der Analyse der Daten mussten gewisse Filterungen vorgenommen werden, insofern, dass manche SA-Verfahren trotz guter Ergebnisse fehlerbehaftet sind, das sie zum Beispiel alle Repliken als negativ bewerten. Für derartige Fälle wurden gefilterte Datensätze erstellt, ebenso auch die Zusammenfassungstabellen. Dies wird in den nachfolgenden Kapiteln aber noch genauer besprochen und betrifft nur vereinzelte Polaritäts-Metriken.

#Namen der Tabellen angeben

#oben noch die F-Wert, Recall, blub False Positives erwähnen (oben ne Tabelle am besten)

### Evaluationsergebnisse pro Lexikon und Polaritäts-Metrik

### Evaluationsergebnisse – SA-Optionen im Vergleich

#NiceToHave

## Diskussion

# Front-End – Visualisierung

## Konzeption

## Entwicklung

## Funktionalität

# Fallbeispiele

# Diskussion

Literaturverzeichnis

Abbasi, A., Hassan, A., & Dhar, M. (2014). Benchmarking Twitter Sentiment Analysis Tools. In LREC (Vol. 14, pp. 26-31).

Appel, O., Chiclana, F., Carter, J., & Fujita, H. (2016). A hybrid approach to the sentiment analysis problem at the sentence level. Knowledge-Based Systems, 108, 110-124.

Asghar, M. Z., Khan, A., Ahmad, S., Qasim, M., & Khan, I. A. (2017). Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. PloS one, 12(2), e0171649.

Chauhan Ashish, P., & Patel, D. K. Sentiment Analysis Using Hybrid Approach: A Survey. Int. Journal of Engineering Research and Applications, 5(1), 73-77.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 2200-2204).

Balahur, A., Steinberger, R., Van Der Goot, E., Pouliquen, B., & Kabadjov, M. (2009). Opinion mining on newspaper quotations. In Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on (Vol. 3, pp. 523-526). IEEE.

Balage Filho, P., & Pardo, T. (2013). NILC\_USP: A Hybrid System for Sentiment Analysis in Twitter Messages. In SemEval@ NAACL-HLT (pp. 568-572).

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., ... & Belyaeva, J. (2013). Sentiment analysis in the news. arXiv preprint arXiv:1309.6202. Retrieved from https://arxiv.org/ftp/arxiv/papers/1309/1309.6202.pdf

Bhatt, A., Patel, A., Chheda, H., & Gawande, K. (2015). Amazon Review Classification and Sentiment Analysis. International Journal of Computer Science and Information Technologies, 6(6), 5107-5110.

Biever, C. (2010). Twitter mood maps reveal emotional states of America. New Scientist, 207(2771), 14.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of computational science, 2(1), 1-8.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, 28(2), 15-21.

Cambria, E., Speer, R., Havasi, C., & Hussain, A. (2010). SenticNet: A Publicly Available Semantic Resource for Opinion Mining. In AAAI fall symposium: commonsense knowledge.

Chin, D., Zappone, A., & Zhao, J. (2016). Analyzing Twitter Sentiment of the 2016 Presidential Candidates. Retrieved from https://web.stanford.edu/~jesszhao/files/twitterSentiment.pdf

Choi, Y., & Cardie, C. (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 793-801). Association for Computational Linguistics.

Cieliebak, M., Dürr, O., & Uzdilli, F. (2013). Potential and Limitations of Commercial Sentiment Detection Tools. In ESSEM@ AI\* IA (pp. 47-58).

Collomb, A., Costea, C., Joyeux, D., Hasan, O., & Brunie, L. (2014). A study and comparison of sentiment analysis methods for reputation evaluation. Rapport de recherche RR-LIRIS-2014-002.

comScore. (2007). Online Consumer-Generated Reviews Have Significant Impact on Offline Purchase Behavior. Retrieved from https://www.comscore.com/Insights/Press-Releases/2007/11/Online-Consumer-Reviews-Impact-Offline-Purchasing-Behavior

Cui, H., Mittal, V., & Datar, M. (2006). Comparative experiments on sentiment classification for online product reviews. In AAAI (Vol. 6, pp. 1265-1270).

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. Management science, 53(9), 1375-1388.

Davidov, D., Tsur, O., & Rappoport, A. (2010a). Enhanced sentiment learning using twitter hashtags and smileys. In Proceedings of the 23rd international conference on computational linguistics: posters (pp. 241-249). Association for Computational Linguistics.

Davidov, D., Tsur, O., & Rappoport, A. (2010b). Semi-supervised recognition of sarcastic sentences in twitter and amazon. In Proceedings of the fourteenth conference on computational natural language learning (pp. 107-116). Association for Computational Linguistics.

D’Andrea, A., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. International Journal of Computer Applications, 125(3), 26-33.

Ding, X., Liu, B., & Yu, P. S. (2008, February). A holistic lexicon-based approach to opinion mining. In Proceedings of the 2008 international conference on web search and data mining (pp. 231-240). ACM.

Donkor, B. (2013). On social Sentiment and Sentiment Analysis. Retrieved from http://brnrd.me/social-sentiment-sentiment-analysis/

Emerson, G. & Declerck, T. (2014). SentiMerge: Combining sentiment lexicons in a Bayesian framework. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing* (pp. 30-38).

Ekman, P., Friesen, W. V., O'sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., ... & Scherer, K. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. Journal of personality and social psychology, 53(4), 712-717.

Elsner, M. (2012). Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 634-644). Association for Computational Linguistics.

Esuli, A., & Sebastiani, F. (2006). Determining Term Subjectivity and Term Orientation for Opinion Mining. In EACL (pp. 193-200).

Esuli, A., & Sebastiani, F. (2007). SentiWordNet: a high-coverage lexical resource for opinion mining. Retrieved from http://nmis.isti.cnr.it/sebastiani/Publications/2007TR02.pdf

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. Journal of Big Data, 2(1), 5.

Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89.

Gindl, S., & Liegl, J. (2008). Evaluation of different sentiment detection methods for polarity classification on web-based reviews. In Proceedings of the 18th European conference on artificial intelligence (pp. 35-43).

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report Stanford, 1-12.

Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). Large-Scale Sentiment Analysis for News and Blogs. ICWSM, 7(21), 219-222.

Horrigan, J. (2008). Online Shopping. Pew Internet and American Life Project Report. Retrieved from http://www.pewinternet.org/files/old-media/Files/Reports/2008/PIP\_Online%20Shopping.pdf.pdf

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.

Im Tan, L., San Phang, W., Chin, K. O., & Anthony, P. (2015). Rule-based sentiment analysis for financial news. In Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on (pp. 1601-1606). IEEE.

Jurafsky, D. & Martin, J. H. (2016). Speech and language processing. Chapter 18. Lexicons for Sentiment and Affect Extraction. Draft of November 7, 2016. Retrieved from https://web.stanford.edu/~jurafsky/slp3/18.pdf

Kaur, A., & Gupta, V. (2013). A survey on sentiment analysis and opinion mining techniques. Journal of Emerging Technologies in Web Intelligence, 5(4), 367-371.

Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. Computational intelligence, 22(2), 110-125.

Khoo, C. S., & Johnkhan, S. B. (2017). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. Journal of Information Science. DOI: 10.1177/0165551517703514.

Khoo, C.S.G., Nourbakhsh, A., & Na, J. C. (2012). Sentiment analysis of online news text: a case study of appraisal theory. Online Information Review, 36(6), 858-878.

Kim, S. M., & Hovy, E. (2004). Determining the sentiment of opinions. In Proceedings of the 20th international conference on Computational Linguistics (p. 1367). Association for Computational Linguistics.

Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg! In Proceedings of the Fifth International Conference on Weblogs and Social Media (pp. 538-541). AAAI Press.

Ku, L., Liang, Y., & Chen, H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. In: Proceedings of AAAI.

Lalji, T., & Deshmukh, S. (2016). Twitter Sentiment Analysis Using Hybrid Approach. International Research Journal of Engineering and Technology, 3(6), 2887-2890.

Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision support systems, 48(2), 354-368.

Liu, B., Hu, M., & Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. In Proceedings of the 14th international conference on World Wide Web (pp. 342-351). ACM.

Liu, B. (2016). Sentiment Analysis. Mining Opinions, Sentiments and Emotions. New York: Cambridge University Press.

Liu, Y., Huang, X., An, A., & Yu, X. (2007). ARSA: a sentiment-aware model for predicting sales performance using blogs. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 607-614). ACM.

Marshall, M. (2009). Sentiment and Accuracy. Retrieved from https://www.lexalytics.com/lexablog/2009/sentiment-and-accuracy

McGlohon, M., Glance, N. S. & Reiter, Z. (2010). Star Quality: Aggregating Reviews to Rank Products and Merchants. In Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2010) (pp. 114-121).

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113.

Melville, P., Gryc, W., & Lawrence, R. D. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1275-1284). ACM.

Mishne, G., & Glance, N. S. (2006). Predicting Movie Sales from Blogger Sentiment. In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs (pp. 155-158).

Missen, M. M. S., Boughanem, M., & Cabanac, G. (2013). Opinion mining: reviewed from word to document level. Social Network Analysis and Mining, 3(1), 107-125.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. Computational Intelligence, 29(3), 436-465.

Mudinas, A., Zhang, D., & Levene, M. (2012). Combining lexicon and learning based approaches for concept-level sentiment analysis. In Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining. ACM.

Musto, C., Semeraro, G., & Polignano, M. (2014). A comparison of lexicon-based approaches for sentiment analysis of microblog posts. Information Filtering and Retrieval, 59.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., & Wilson, T. (2013). Semeval-2013 task 2: Sentiment analysis in twitter. In Second Joint Conference on Lexical and Computational Semantics (\* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013) (Vol. 2, pp. 312-320).

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903. Retrieved from http://www2.imm.dtu.dk/pubdb/views/edoc\_download.php/6006/pdf/imm6006.pdf

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (pp. 122-129).

Ogneva, M. (2010). How companies can use sentiment analysis to improve their business. Retrieved from http://mashable.com/2010/04/19/sentiment-analysis/#cyaOLjC5C5q0

Pak, A., & Paroubek, P. (2010). Twitter based system: Using Twitter for disambiguating sentiment ambiguous adjectives. In Proceedings of the 5th International Workshop on Semantic Evaluation (pp. 436-439). Association for Computational Linguistics.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd annual meeting on association for computational linguistics (pp. 115-124). Association for Computational Linguistics.

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

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic inquiry and word count: LIWC [Computer software]. Austin, TX: liwc. net.

Polanyi, L., & Zaenen, A. (2006). Contextual Valence Shifters. Computing Attitude and Affect in Text, 20, 1-10.

Qiu, G., He, X., Zhang, F., Shi, Y., Bu, J., & Chen, C. (2010). DASA: dissatisfaction-oriented advertising based on sentiment analysis. Expert Systems with Applications, 37(9), 6182-6191.

Rainie, L. & Horrigon, J. (2007). Election 2006 Online. Pew Internet and American Life Project Report. Retrieved from http://www.pewinternet.org/2007/01/17/election-2006-online/

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge-Based Systems, 89, 14-46.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In Proceedings of the ACL student research workshop (pp. 43-48). Association for Computational Linguistics.

Remus, R., Quasthoff, U. & Heyer, G. (2010). SentiWS-A Publicly Available German-language Resource for Sentiment Analysis. In LREC (pp. 1168-1171).

Scaffidi, C., Bierhoff, K., Chang, E., Felker, M., Ng, H., & Jin, C. (2007). Red Opal: product-feature scoring from reviews. In Proceedings of the 8th ACM conference on Electronic commerce (pp. 182-191). ACM.

Singh, V. K., Piryani, R., Uddin, A., & Waila, P. (2013). Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013 International Multi-Conference on (pp. 712-717). IEEE.

Sommar, F., & Wielondek, M. (2015). Combining Lexicon-and Learning-based Approaches for Improved Performance and Convenience in Sentiment Classification. Retrieved from http://www.diva-portal.org/smash/get/diva2:811021/fulltext01.pdf

Tan, L. K. W., Na, J. C., Theng, Y. L., & Chang, K. (2012). Phrase-level sentiment polarity classification using rule-based typed dependencies and additional complex phrases consideration. Journal of Computer Science and Technology, 27(3), 650-666.

Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558.

Tsai, A. C. R., Wu, C. E., Tsai, R. T. H., & Hsu, J. Y. J. (2013). Building a concept-level sentiment dictionary based on commonsense knowledge. IEEE Intelligent Systems, 28(2), 22-30.

Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web. Data Mining and Knowledge Discovery, 24(3), 478-514.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (pp. 178-185).

Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey. International Journal of Advanced Research in Computer Science and Software Engineering, 2(6), 282-292.

Võ, M. L., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin affective word list reloaded (BAWL-R). Behavior research methods, 41(2), 534-538.

Võ, M. L., Jacobs, A. M., & Conrad, M. (2006). Cross-validating the Berlin affective word list. Behavior research methods, 38(4), 606-609.

Wanner, F., Rohrdantz, C., Mansmann, F., Stoffel, A., Oelke, D., Krstajic, M., ... & Atkinson, M. (2009). Large-scale comparative sentiment analysis of news articles. In IEEE Information Visualization Conference : InfoVis 2009.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on human language technology and empirical methods in natural language processing (pp. 347-354). Association for Computational Linguistics.

Xu, K., Liao, S. S., Li, J., & Song, Y. (2011). Mining comparative opinions from customer reviews for Competitive Intelligence. Decision support systems, 50(4), 743-754.

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through twitter “I hope it is not as bad as I fear”. Procedia-Social and Behavioral Sciences, 26, 55-62.

Anhang A – B5T-Fragebogen