Universität Regensburg

„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: 1441937
Erstgutachter: Prof. Dr. Christian Wolff
Zweitgutachter: Dr. Manuel Burghardt
Laufendes Semester: WS 2017/2018
Abgegeben am: 09.10.2017

Inhalt

1 Einleitung 12

2 Related Work 14

2.1 Sentiment Analysis – Grundlagen 14

2.2 Lexikon-basierte Ansätze 22

2.3 Deutschsprachige Sentiment-Lexika 29

2.3.1 SentimentWortschatz – SentiWS 30

2.3.2 Berlin Affective Word List – Reloaded (BAWL-R) 31

2.3.3 NRC Word-Emotion Association Lexicon (NRC) 33

2.3.4 Clematide-Dictionary (CD) 35

2.3.5 German Polarity Clues (GPC) 37

2.3.6 Sonstige SA-Lexika 39

2.4 Sentiment Analysis in der Literaturwissenschaft 41

3 Forschungsfrage und -Agenda 53

4 Dramen-Korpus 59

5 Back End: Sentiment Analysis 64

5.1 Konzeption 64

5.2 Vorverarbeitung der Dramen 65

5.2.1 Allgemeine Vorverarbeitung 65

5.2.2 Sprachverarbeitung – Lemmatisierung 67

5.2.2.1 Idee 67

5.2.2.2 Entwicklung 69

5.2.3 Stoppwortlisten 70

5.2.3.1 Idee 70

5.2.3.2 Entwicklung 73

5.3 Lexika-Verarbeitung 73

5.3.1 Grundsätzliche Verarbeitung und Lexikon-Auswahl 73

5.3.2 Lexika-Kombination 76

5.3.2.1 Idee 76

5.3.2.2 Entwicklung 78

5.3.3 DTA – Erweiterung 79

5.3.3.1 Idee 79

5.3.3.2 Entwicklung 81

5.3.4 Sprachverarbeitung – Lemmatisierung 82

5.3.4.1 Idee 82

5.3.4.2 Entwicklung 84

5.4 Vorverarbeitung – Überblick 84

5.5 Sentiment Analysis 86

5.5.1 Konzeption ... 86
5.5.2 SA-Metriken ... 87
5.5.3 Entwicklung ... 93
6 Vokabular-basierte Evaluation ... 96
6.1 Idee und Vorgehen ... 96
6.2 Entwicklung ... 98
6.3 Ergebnisse ... 100
6.4 Fazit ... 104
7 Erstellung des Gold-Standard-Korpus ... 105
7.1 Test-Korpus-Erstellung ... 106
7.1.1 Idee und Vorgehen ... 106
7.1.2 Entwicklung ... 109
7.1.3 Ergebnisse ... 110
7.2 Test-Korpus-Annotation ... 111
7.2.1 Idee und Vorgehen ... 111
7.2.1.1 Annotationsschema ... 111
7.2.1.2 Durchführung und Stichprobe ... 114
7.2.1.3 Fragebogen ... 117
7.2.2 Ergebnisse ... 118
7.2.2.1 Datenaufbereitung ... 118
7.2.2.2 Entwicklung und produzierte Datenstrukturen ... 119
7.2.2.3 Sentiment-Verteilungen ... 120
7.2.2.4 Annotatoren-Übereinstimmung ... 129
7.2.2.5 Mehrheitsannotationen ... 139
7.2.2.6 Fragebogen-Auswertung ... 143
7.3 Diskussion ... 145
8 SA-Evaluation ... 148
8.1 Vorgehen ... 148
8.2 Entwicklung ... 151
8.3 Ergebnisse ... 153
8.3.1 Datenaufbereitung ... 153
8.3.2 Benchmark ... 155
8.3.3 Evaluationsergebnisse pro Lexikon und Polaritäts-Metrik ... 156
8.3.3.1 Sentiment Wortschatz – SentiWS ... 157
8.3.3.2 Berlin Affective Word List – Reloaded (BAWL-R) ... 161
8.3.3.3 NRC Word-Emotion Association Lexikon (NRC) ... 164
8.3.3.4 Clematide-Dictionary (CD) ... 166
8.3.3.5 German Polarity Clues (GPC) ... 169
8.3.3.6 Kombiniertes Lexikon ... 173

8.3.4 Evaluationsergebnisse – Metriken im Vergleich ... 177
8.4 Diskussion und Fazit ... 180
9 Front-End – Visualisierung ... 184
9.1 Idee und Motivation ... 184
9.2 Verwendete Metriken ... 185
9.3 Entwicklung ... 191
9.4 Funktionalität ... 192
9.4.1 Allgemeines Design ... 193
9.4.2 Header ... 193
9.4.3 Dramen-Auswahl ... 194
9.4.4 Strukturelle Analyse ... 194
9.4.4.1 Sentiments im ganzen Drama ... 195
9.4.4.2 Kreisdiagramm – Sentiment-Anteile im Drama ... 196
9.4.4.3 Verlaufsdiagramm – Sentiments im Drama pro Akt ... 198
9.4.4.4 Kreisdiagramm – Sentiment-Anteile pro Akt ... 199
9.4.4.5 Verlaufsdiagramm – Sentiments in Szenen pro Akt ... 199
9.4.4.6 Kreisdiagramm – Sentiment-Anteile pro Szene ... 200
9.4.4.7 Verlaufsdiagramm – Szenen im Dramenverlauf ... 201
9.4.4.8 Verlaufsdiagramm – Repliken ... 202
9.4.5 Sprecher-Analyse ... 204
9.4.5.1 Sentiments im ganzen Drama ... 205
9.4.5.2 Verlaufsdiagramm – Sprecher-Sentiments pro Akt ... 206
9.4.5.3 Verlaufsdiagramm – Sprecher-Sentiments in Szenen pro Akt ... 207
9.4.5.4 Verlaufsdiagramm – Szenen im Dramenverlauf pro Sprecher ... 208
9.4.5.5 Verlaufsdiagramm pro Sprecher – Repliken ... 210
9.4.5.6 Statischer Sprechervergleich ... 212
9.4.5.7 Kreisdiagramm - Sentiment-Anteile von Sprecher ... 214
9.4.6 Sprecher-Beziehungen (Charakter-zu-Charakter-Sentiment) ... 215
9.4.6.1 Sprecher-Beziehungs-Sentiments im Drama ... 216
9.4.6.2 Sprecher-Beziehungs-Sentiments pro Akt ... 217
9.4.6.3 Sprecher-Beziehungs-Sentiments pro Szene ... 218
9.4.6.4 Kreisdiagramm - Sentiment-Anteile von Sprecherbeziehungen ... 220
10 Fallbeispiele ... 221
10.1 Fallbeispiel 1: Polaritäten im Aktverlauf ... 222
10.2 Fallbeispiel 2: Marinelli in Emilia Galotti ... 224
11 Diskussion und Ausblick ... 227
Literaturverzeichnis ... 234
Anhang – DVD ... 245
Plagiatserklärung ... 247

Abbildungen

Abbildung 1: Ausschnitt SentiWS 30
Abbildung 2: Ausschnitt BAWL-R 32
Abbildung 3: Ausschnitt CD 36
Abbildung 4: Ausschnitt GPC 38
Abbildung 5: Ausschnitt TextGrid-Drama 61
Abbildung 6: Beispiel DTA-Format 80
Abbildung 7: Ausschnitt Gesamt-Korpus Vokabular lemmatisiert 99
Abbildung 8: Ausschnitt Vokabular-basierte Evaluationsdatei 100
Abbildung 9: Vokabular-basierte Evaluation – Balkendiagramm – Lexikonvergleich 101
Abbildung 10: Vokabular-basierte Evaluation – Balkendiagramm – Lexikonvergleich mit DTA-Erweiterung 102
Abbildung 11: Vokabular-basierte Evaluation – Balkendiagramm – Lexikonvergleich mit Lemmatisierung 103
Abbildung 12: Ausschnitt Beispielreplik des Test-Korpus 110
Abbildung 13: Beispiel-Annotation 115
Abbildung 14: Sentiment-Verteilung Polarität Standard 121
Abbildung 15: Sentiment-Verteilung Polarität Reduziert (vierfach) 122
Abbildung 16: Sentiment-Verteilung Polarität Dichotom 123
Abbildung 17: Balkendiagramm Polarität Reduziert (vierfach) * Polarität Dichotom 124
Abbildung 18: Mittelwert von Länge verteilt auf Polarität Reduziert (vierfach) 125
Abbildung 19: Balkendiagramm – Häufigkeitsverteilung Emotionskategorien 126
Abbildung 20: Häufigkeitsverteilung – Emotion vorhanden 127
Abbildung 21: Häufigkeitsverteilungen – Emotionskategorien 128
Abbildung 22: Mittelwerte von Länge für Emotion vorhanden 129
Abbildung 23: K-Alpha für Polaritätsvariablen 132
Abbildung 24: Prozentuale Durchschnittsübereinstimmung für Polaritätsvariablen 133
Abbildung 25: Prozentuale Durchschnittsübereinstimmung für Polarität Dichotom für alle Dramen 134
Abbildung 26: Häufigkeitsverteilung Mehrheitsgruppen Polarität Standard 135
Abbildung 27: Häufigkeitsverteilung Mehrheitsgruppen Polarität reduziert (vierfach) 135
Abbildung 28: Häufigkeitsverteilung Mehrheitsgruppen Polarität Dichotom 136
Abbildung 29: Prozentuale Durchschnittsübereinstimmung Emotionskategorien 137
Abbildung 30: Häufigkeitsverteilung Mehrheitsentscheidungen Polarität Standard 139

Abbildung 31: Häufigkeitsverteilung Mehrheitsentscheidungen Polarität reduziert (vierfach) 140
Abbildung 32: Häufigkeitsverteilung Mehrheitsentscheidungen Polarität Dichotom 142
Abbildung 33: Häufigkeitsverteilung Mehrheitsentscheidungen Emotionskategorien 143
Abbildung 34: Ausschnitt Ergebnistabellen Evaluationsstudie 154
Abbildung 35: Accuracy im Vergleich pro Metrik 178
Abbildung 36: Header 194
Abbildung 37: Dramen-Auswahl 194
Abbildung 38: Visualisierung – Sentiments im ganzen Drama (Polarität gewichtet) 195
Abbildung 39: Visualisierung – Sentiments im ganzen Drama (Emotionen) 196
Abbildung 40: Visualisierung – Sentiment-Anteile im Drama I 197
Abbildung 41: Visualisierung – Sentiment-Anteile im Drama II 197
Abbildung 42: Visualisierung – Verlaufsdiagramm – Sentiments im Drama pro Akt 198
Abbildung 43: Visualisierung – Kreisdiagramm – Sentiment-Anteile pro Akt 199
Abbildung 44: Visualisierung – Ausschnitt Verlaufsdiagramm – Sentiments in Szenen pro Akt 200
Abbildung 45: Visualisierung – Kreisdiagramm – Sentiment-Anteile pro Szene 201
Abbildung 46: Visualisierung – Verlaufsdiagramm – Szenen im Dramenverlauf 202
Abbildung 47: Visualisierung – Verlaufsdiagramm – Repliken 203
Abbildung 48: Visualisierung – Verlaufsdiagramm – Repliken (mit Zoom-Funktion) 204
Abbildung 49: Visualisierung – Sentiments im ganzen Drama (Sprecher) 205
Abbildung 50: Visualisierung – Verlaufsdiagramm – Sprecher-Sentiments pro Akt 206
Abbildung 51: Visualisierung – Ausschnitt – Sprecher-Sentiments in Szenen pro Akt 208
Abbildung 52: Visualisierung – Verlaufsdiagramm – Szenen im Dramenverlauf pro Sprecher 209
Abbildung 53: Visualisierung – Verlaufsdiagramm pro Sprecher – Repliken 211
Abbildung 54: Visualisierung - Sprechervergleich 213
Abbildung 55: Visualisierung – Kreisdiagramm – Sentiment-Anteile von Sprecher 215
Abbildung 56: Visualisierung – Sprecher-Beziehungs-Sentiments im Drama 217
Abbildung 57: Visualisierung – Sprecher-Beziehungs-Sentiments pro Akt 218
Abbildung 58: Visualisierung – Sprecher-Beziehungs-Sentiments pro Szene 219
Abbildung 59: Kreisdiagramm – Sentiment-Anteile von Sprecherbeziehungen 221
Abbildung 60: Polaritäten im Akt-Verlauf bei „Der Freigeist“ 223
Abbildung 61: Polaritäten im Aktverlauf bei „Emilia Galotti“ 223
Abbildung 62: Polarität für die Figur Marinelli in Emilia Galotti 224
Abbildung 63: Sprecher-Vergleich für Polaritäten in Emilia Galotti 225

Abbildung 64: Sprecher-Beziehungspolaritäten aus Sicht der Figur Claudia..... 226
Abbildung 65: Sprecherbeziehungspolaritäten aus Sicht der Figur Marinelli..... 226

Tabellen

Tabelle 1: Forschungsagenda 58
Tabelle 2: Korpus – Dramen-Statistiken 63
Tabelle 3: Korpus – Gesamt-Statistiken 64
Tabelle 4: Verarbeitungs-Optionen – DTA-Erweiterung 85
Tabelle 5: Verarbeitungs-Optionen – Lemmatisierer 85
Tabelle 6: Verarbeitungs-Optionen – Lemmatisierungstyp 85
Tabelle 7: Verarbeitungs-Optionen – Stoppwortlisten 85
Tabelle 8: Verarbeitungs-Optionen – Groß- und Kleinschreibung 86
Tabelle 9: Sentiment-Metriken SentiWS 89
Tabelle 10: Sentiment-Metriken BAWL-R 89
Tabelle 11: Sentiment-Metriken NRC 90
Tabelle 12: Sentiment-Metriken CD 90
Tabelle 13: Sentiment-Metriken GPC 91
Tabelle 14: Sentiment-Metriken Kombiniertes Lexikon 91
Tabelle 15: Sentiment-Metriken Normalisierungen 93
Tabelle 16: Test-Korpus Statistiken 110
Tabelle 17: Annotationsschema Polarität 113
Tabelle 18: Annotationsschema Emotionen 114
Tabelle 19: Kreuztabelle Polarität Reduziert (vierfach)*Polorität Dichotom 123
Tabelle 20: Verteilung Emotionskategorien 126
Tabelle 21: Mehrheitsverteilungen Emotionskategorien 138
Tabelle 22: Fragebogen – Schwierigkeit und Sicherheit – Statistiken 144
Tabelle 23: Fragebogen – Zeit in Minuten – Statistiken 145
Tabelle 24: Kreuztabelle – Prädiktionsmöglichkeiten 150
Tabelle 25: Übersicht Evaluationsmaße polaritySentiWS 157
Tabelle 26: Ausschnitt Evaluationsmaße Verfahrenskombinationen polaritySentiWS 158
Tabelle 27: Übersicht Evaluationsmaße polaritySentiWSDichotom 160
Tabelle 28: Ausschnitt Evaluationsmaße Verfahrenskombinationen polaritySentiWSDichotom 160
Tabelle 29: Übersicht Evaluationsmaße emotion 162
Tabelle 30: Ausschnitt Evaluationsmaße Verfahrenskombinationen emotion 162
Tabelle 31: Übersicht Evaluationsmaße polarityBawlDichotom 163
Tabelle 32: Ausschnitt Evaluationsmaße Verfahrenskombinationen polarityBawlDichotom 164

Tabelle 33: Übersicht Evaluationsmaße polarityNrc 165
Tabelle 34: Ausschnitt Evaluationsmaße Verfahrenskombinationen polarityNrc 165
Tabelle 35: Übersicht Evaluationsmaße polarityCd 166
Tabelle 36: Ausschnitt Evaluationsmaße Verfahrenskombinationen polarityCd 167
Tabelle 37: Übersicht Evaluationsmaße polarityCdDichotom 168
Tabelle 38: Ausschnitt Evaluationsmaße Verfahrenskombinationen polarityCDDichotom . 169
Tabelle 39: Ausschnitt Evaluationsmaße Verfahrenskombinationen polarityGpc 170
Tabelle 40: Prädiktionstabelle polarityGpc 170
Tabelle 41: Übersicht Evaluationsmaße polarityGpc 172
Tabelle 42: Ausschnitt Evaluationsmaße Verfahrenskombinationen polarityGpc (gefiltert) 172
Tabelle 43: Übersicht Evaluationsmaße polarityCombined 174
Tabelle 44: Ausschnitt Evaluationsmaße Verfahrenskombinationen polarityCombined 174
Tabelle 45: Übersicht Evaluationsmaße clearlyPolarityCombined 176
Tabelle 46: Ausschnitt Evaluationsmaße Verfahrenskombinationen clearlyPolarityCombined 176
Tabelle 47: Back-End-Metriken übertragen auf Front-End-Namen 188
Tabelle 48: Back-End-Normalisierungen übertragen auf Front-End-Namen 189
Tabelle 49: Sentiment-Gruppen im Front-End 190

Zusammenfassung

In der vorliegenden Masterarbeit wird ein mehrteiliges Projekt vorgestellt, das den Einsatz von Sentiment Analysis (SA) in der quantitativen Dramenanalyse exploriert. Als beispielhafter Untersuchungsgegenstand wird ein Korpus von 11 Dramen des Schriftstellers Gotthold Ephraim Lessing (1729 – 1782) verwendet. Die Arbeit stellt eine Erweiterung eines bestehenden Tools zur quantitativen Dramenanalyse (Katharsis) um eine SA-Komponente dar.

Es wurden Python-Programme zur Durchführung der SA entwickelt. Als zentraler SA-Ansatz wird mangels annotierter Trainings-Korpora ein Lexikon-basierter Ansatz gewählt. Um ein optimiertes SA-Verfahren zu identifizieren, werden mehrere Optionen und Herangehensweisen für die SA implementiert und auf ihre Leistung für den spezifischen Anwendungsfall untersucht. Es werden fünf der bekanntesten deutschsprachigen SA-Lexika implementiert sowie eine kombinierte Gesamtversion dieser erstellt. Als weitere Optionen wird der Einfluss einer Lexikonerweiterung mit historischen linguistischen Varianten, von Lemmatisierung über zwei Lemmatisierer und drei Lemmatisierungsarten, von drei verschiedenen Stoppwortlisten und der Beachtung von Groß- und Kleinschreibung implementiert und untersucht. Es werden für alle kombinatorischen Möglichkeiten von Lexika und Optionen verschiedene Sentiment-Metriken auf verschiedenen Ebenen berechnet. Als Ebenen des Dramas werden Sentiment-Metriken für die strukturelle Ebene (Drama, Akt, Szene, Replik), die Sprecher-Ebene (pro Drama, Akt, Szene, Replik) und für Sprecherbeziehungen (pro Drama, Akt, Szene, Replik) kalkuliert. Es werden unterschiedliche Metriken für die Polarität (positiv, negativ) und 8 Emotionskategorien auf diesen Ebenen berechnet.

Es werden mehrere Evaluationsverfahren durchgeführt. In einer ersten informellen Evaluation wird der Anteil der Wörter der Lexika in Zusammenhang mit den genannten Optionen am Vokabular des Korpus untersucht und diskutiert. Zur Ausführung einer systematischen Evaluation wird ein Gold-Standard von annotierten Repliken erstellt. In einer Annotationsstudie beurteilen 5 Teilnehmer einen repräsentativen Korpus von 200 Repliken bezüglich Polarität und Emotionen. In einem anschließenden Fragebogen konnten Einsichten zu Probleme und Schwierigkeiten bei der Annotation erhoben werden. Die Ergebnisse der Annotation werden statistisch ausgewertet und

hinsichtlich Annotationsverhalten untersucht. Als Hauptergebnisse stellt man einen grundsätzlich geringeren Übereinstimmungsgrad als bei anderen Untersuchungsgegenständen in der SA fest. Auffällig ist auch eine starke Ungleichverteilung der Polaritäten im Korpus. Es werden deutlich mehr Repliken als negativ denn als positiv wahrgenommen. Das finale Evaluations-Korpus (Gold Standard, GS) besteht aus 139 negativen und 61 positiven Repliken basierend auf der Mehrheitsentscheidung der Annotatoren.

Über ein in Python entwickeltes Evaluationsframework wurde systematisch die SA-Leistung aller Lexika und Methoden hinsichtlich der Prädiktion der Polarität einer Replik untersucht. Verschieden Evaluations-Metriken wurden zur differenzierten Analyse und Diskussion aller Ansätze berechnet. Es können Erkennungsraten von bis zu 70% festgestellt werden. Unter Analyse aller Evaluationsergebnisse wird das leistungsstärkste Verfahren bestimmt. Es setzt sich aus der Methoden-Kombination des Lexikons SentiWS, erweitert durch historische linguistische Varianten, mit einer Lemmatisierung auf Text- und Lexikon-Ebene über den pattern-Lemmatisierer, ohne Stoppwortiste und unter Beachtung von Groß- und Kleinschreibung im letzten Abgleichschritt, zusammen.

Für das als am besten identifizierte Verfahren wird ein Front-End zur Visualisierung der SA-Metriken als Web-Anwendung implementiert. Es stehen interaktive Visualisierungen für Polaritäten und Emotionskategorien zur Verfügung. Es können Verteilungen und Verläufe auf Dramen-, Akt-, Szenen-, Replik-, Sprecher- und Sprecherbeziehungs (je pro Drama, Akt, Szene, Replik) exploriert werden. Der mögliche Einsatz in der Dramenanalyse wird anhand vereinzelter Fallbeispiele beschrieben. Abschließend werden die Ergebnisse des Gesamtprojekts im Kontext der Forschung diskutiert und mögliche Anknüpfungspunkte besprochen.

1 Einleitung

In der vorliegenden Arbeit werden zwei Strömungen aktueller Forschung der Digital Humanities, die Sentiment Analysis (SA) und die quantitative Dramenanalyse zusammengeführt. Als SA wird ein Studienfeld definiert, das sich mit der computergestützten Prädiktion von Gefühlen, Meinungen und Emotionen in geschriebenen Texten (Liu, 2016) befasst. Die wichtigsten Anwendungsgebiete der SA sind bislang vor allem Online-Reviews (Cui, Mittal & Datar, 2006; McGlohan, Glance & Reiter, 2010) sowie Social Media (Pak & Paroubek, 2010; Kouloumpis, Wilson & Moore, 2011). Metastudien zur SA legen einen deutlichen Mangel an Projekten bezüglich literarischer Texte nah (Tsytsarau & Palpanas, 2012; Ravi & Ravi, 2015).

Die quantitative Dramenanalyse ist ein Teilbereich der quantitativen Literaturwissenschaft, die versucht die hermeneutische Arbeitsweise in der Literaturwissenschaft mit mathematisch-quantitativen Methoden zu unterstützen und zu erweitern (Fucks & Lauter, 1965). Speziell für die quantitative Dramenanalyse entwickelte Solomon (1971; 1973) ein mathematisches Modell, dessen Grundkonzepte gewinnbringend in der quantitativen Dramenanalyse eingesetzt werden konnten (Ilsemann, 2005; 2008). Die SA wird als mathematisch-linguistische Methode für den Einsatz auf literarischen Texten wie Dramen in der vorliegenden Arbeit als zugehörig zum Forschungsgebiet der quantitativen Dramenanalyse betrachtet.

Der Einsatz von SA auf literarischen Texten und Dramen ist bislang selten, obschon Emotionen und Gefühle zentrale Komponenten von Literatur und literaturwissenschaftlicher Interpretation sind (Alt, 1994, S. 191-210; Fick, 2000, S. 334-335; Winko, 2003; Mellmann, 2007; Mohammad, 2011; Nalisnick & Baird, 2013). Erste Projekte und Studien untersuchen beispielsweise literarische Gattungen wie Märchen (Alm & Sproat, 2005a; Alm et al., 2005; Volkova et al., 2010; Mohammad, 2011), Romanen (Kakkonen & Kakkonen, 2011; Elsner, 2012; Jannidis et al., 2016; Klinger, Suliya & Reiter, 2016). Im Bereich der Dramenanalyse wurden Emotionsverteilungen (Mohammad, 2011) und Figurenbeziehungen (Nalisnick & Baird, 2013) in Shakespeare-Dramen mittels SA analysiert. Eine Anwendung von SA auf deutschsprachigen Dramen ist bislang nicht bekannt.

Im vorliegenden Projekt wird der Einsatz von SA in der Dramenanalyse auf mehreren Ebenen und über mehrere Arbeitspakete differenziert untersucht, um sich mit der Frage auseinanderzusetzen, ob und inwiefern die SA in der Literaturwissenschaft gewinnbringend genutzt werden kann. Es werden zentrale Probleme der bisherigen Forschung, wie der Mangel annotierter Test-Korpora und die fehlende systematische Evaluation verschiedener Herangehensweisen aufgegriffen um mit der vorliegenden Arbeit einen bedeutenden Mehrwert für die bisherige Forschung und die Literaturwissenschaft zu liefern. Das Projekt ist dabei in großen Teilen eine Erweiterung eines bestehenden Tools zur quantitativen Dramenanalyse (Burghardt et al., 2016) um eine SA-Komponente.

Im nachfolgenden Kapitel 2 wird die relevante Literatur zum Thema SA mit einem Fokus auf die in dieser Arbeit verwendeten Lexikon-basierten Verfahren zusammengefasst. Es werden im Deutschen verfügbare Sentiment-Lexika im Detail beschrieben und Projekte zum Themenbereich der SA auf literarischen Texten präsentiert. In Kapitel 3 wird die zentrale Motivation der Arbeit erläutert, eine allgemeine Forschungsfrage zur Orientierung formuliert und die einzelnen Schritte der Forschungsagenda zusammengefasst. In Kapitel 4 wird der im vorliegenden Projekt verwendete Korpus von Lessing-Dramen beschrieben. Kapitel 5 befasst sich mit der konkreten Entwicklung und Konzeption eines Back-Ends zur Umsetzung verschiedener SA-Verfahren und Kalkulation unterschiedlicher SA-Metriken auf zahlreichen Dramen-Ebenen. In Kapitel 6 wird eine erste informelle Evaluation basierend auf Wortabgleich präsentiert (Vokabular-basierte Evaluation). Kapitel 7 nun schildert eine Teilstudie des Projekts, bei der durch Annotation von 5 Teilnehmern ein Gold Standard (GS) erstellt wurde. Das Annotationsverhalten wird statistisch untersucht und in den Kontext bisheriger Annotation-Forschung für literarische Texte gestellt. Ferner wird die Zusammensetzung des finalen GS analysiert. In Kapitel 8 werden die entwickelten SA-Verfahren systematisch über ein eigens implementiertes Framework evaluiert und verglichen, um das im Vergleich optimalste Verfahren zu identifizieren. Die Resultate werden diskutiert und interpretiert. In Kapitel 9 wird eine auf Basis der SA-Verfahren entwickelte Web-Anwendung zur Visualisierung und Exploration von SA-Metriken für die Dramenanalyse vorgestellt. Anhand vereinzelter Fallbeispiele wird in Kapitel 10 der potentielle Einsatz in der Literaturwissenschaft analysiert. In der abschließenden Gesamt-Diskussion in Kapitel 11 werden

die zentralen Resultate zusammengefasst und im Kontext der Forschung besprochen. Es werden Grenzen des Projekts und mögliche Anknüpfungspunkte besprochen. Separate Diskussionskapitel für zwei größere Teilpakete, die Annotationsstudie und die Evaluation, findet man in Kapitel 7.3 und Kapitel 8.4. In einem Anhang auf DVD findet man alle Programm-Dateien, Auswertungen sowie Videos zur Nutzung des Fron-Ends. Da das Projekt auf ein bestehendes Tool aufbaut wird in den nachfolgenden Kapiteln stets exakt angegeben, welche Dateien neu entwickelt und angepasst oder erweitert wurden. Alle Auswertungen über Tabellen oder SPSS-Dateien wurden generell erst im vorliegenden Projekte durchgeführt.

2 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.

2.1 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, das 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 natürlich-sprachlichen 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 Autoritäten 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. Zuletzt sei noch die Analyse von Börsen- und Marktnachrichten genannt, 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 2.2) 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 Klassifikationsansätze 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 nume-

rische 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 den Häufigkeiten einzelner Wörter und n-Grammen als Merkmalsvektoren eines Dokuments. Es werden die drei Standard-Algorithmen Naive Bayes, Maximum-Entropie-Methodik und Support Vector Machines 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 Erzeugung 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 sich die Mehrzahl der Forschung auf Produkt- oder Film-Reviews fokussiert und es für diese Bereiche sehr einfach ist annotierte Trainingskorpora zu akquirieren, um themenspezifisch 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 8 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 wurden 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 visualisieren. Damit soll die Marktanalyse und das Risiko-Management eines Unternehmens unterstützt werden. 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).

2.2 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 (Taboada 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 vorherrschende Verwendungsweise in der SA ist (siehe Kapitel 2.1). 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 englisch-

sprachige 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 genannt (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 & Inkpen, 2006; Tsytsarau & Palpanas, 2012). Im erstgenannten Beispiel hätte der Satz eine negative Ausprägung von fünf, 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 insgesamt ein stärkeres Gewicht haben als die häufigeren 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 2.1 geschildert kann man den Einsatz von Sentiment-Lexika verfeinern und präzisieren. Dazu gehören 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 und Sarkasmus auf linguistischer Ebene (Carvalho et al., 2009; Reyes, Rosso & Veale 2013) und den Umgang mit dieser in der SA (Polanyi & Zaenen, 2006; Maynard & Greenwood, 2014). Aufgrund der an-

spruchsvollen 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 (Kennedy & Inkpen, 2006; Taboada et al., 2011). 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 5.2.2 und 5.3.4 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 bestimmte Wortarten für die SA genutzt wird (Denecke, 2008) oder dass Wortambiguitäten über die POS aufgelöst werden (Taboada et al, 2011). Beim Einsatz von Lexikon-basierten Verfahren kann die Vorverarbeitung des zu analysierenden Textes als Verarbeitungskette mit mehreren Schritten betrachtet werden (Denecke, 2008; Zhang et al., 2011; Musto et al., 2014). 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; z.B. bei Waltinger, 2010; oder Momtazi, 2012). 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 Sy-

nonymen 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 domänenspezifische 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 Methoden 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 2.1 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, das 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 und 4912 negative. Es wurde durch über mehrere Methoden erstellt, unter anderem die Kombination bestehender Datensätze und die manuelle 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 (2012) 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). 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 (Mohammad & Turney, 2010), die Mehrzahl der aktuelleren SA-Lexika wurde jedoch maschinell erstellt und nicht vollständig manuell korrigiert, 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 zwingend 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 Metaphern geschehen.

Bereits in Kapitel 2.1 wurden Evaluationsergebnisse beim Vergleich von ML- und Lexikon-Ansätzen beschrieben. Eine Zusammenfassung von Evaluationsergebnissen in der Forschung zu unterschiedlichen SA-Ansätzen findet man in Tsytsarau und Palpanas (2012). 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, 2012).

Wurden in Studien ML-Algorithmen mit SA-Lexika unter den gleichen Bedingungen (gleicher Evaluationskorpus, gleiches Evaluationsframework) verglichen, schneiden ML-Algorithmen besser ab (Giendl & Liegl, 2008; Pang et al., 2002). Als konkretes Beispiele mit guten Erkennungsraten 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. Aus 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 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.

2.3 Deutschsprachige Sentiment-Lexika

In diesem Abschnitt 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

(Remus, Quasthoff & Heyer, 2010). Für jedes Lexikon werden 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.

2.3.1 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:

```
327  Kritik|NN   -0.5308 Kritiken
328  Kritiker|NN -0.6494 Kritikers,Kritikern
329  Kränkung|NN -0.0048 Kränkungen
330  Krüppel|NN  -0.3203 Krüppeln,Krüppels
331  Kurseinbruch|NN -0.0048 Kurseinbruchs,Kurseinbrüche,Kurseinbruches,Kurseinbrüchen
332  Kälte|NN    -0.0048
333  Kündigung|NN    -0.0048 Kündigungen
334  Kürzung|NN  -0.341  Kürzungen
335  Langeweile|NN   -0.0377
336  Langweiler|NN   -0.0443 Langweilern,Langweilers
337  Last|NN -0.0048 Lasten
338  Launenhaftigkeit|NN -0.0048
339  Lebensgefahr|NN -0.0048 Lebensgefahren
```

Abbildung 1: Ausschnitt SentiWS

Das Lexikon wurde auf Basis von drei Quellen generiert. Zunächst wurden die positiven und negativen Kategorien des bereits angesprochenen (siehe Kapitel 2.2) 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 Me-

thoden 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 zusätzlich manuell auf ihre Eignung als SBW überprüft wurden. Die einzelnen Polaritätsgewichte wurden sodann mittels der semantischen Nähe zu Ausgangswörtern (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 menschliche Beurteiler 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 8.1). Insgesamt erbringt SentiWS also eine sehr gute Leistung 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.

2.3.2 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 (Russel, 1980). 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 hinsichtlich 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 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.

	WORD	WORD_LOW	WORD_CLAS	EMO_MEAN	EMO_STD	AROUSAL_MEA
2	AAL	aal	N	-0,5	0,70710678	2,380952381
3	AAS	aas	N	-2,1	1,10050493	2,631578947
4	ABART	abart	N	-1,6	0,6992059	3,277777778
5	ABBAU	abbau	N	-1	1,1697953	3
6	ABBAUEN	abbauen	V	-0,8	0,92	2,105263158
7	ABBILD	abbild	N	-0,2	0,63245553	2,105263158
8	ABBRUCH	abbruch	N	-0,7	1,15950181	2,904761905
9	ABDANKEN	abdanken	V	-0,4	0,84	2,666666667
10	ABDRUCK	abdruck	N	-0,1	0,31622777	2,235294118
11	ABEND	abend	N	1,65	0,933302	1,833333333

Abbildung 2: Ausschnitt BAWL-R

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 Informationen wurden gesammelt und sind Bestandteil des Lexikons, spielen jedoch für die weitere SA keine Rolle.

2.3.3 NRC Word-Emotion Association Lexicon (NRC)

Das NRC Word-Emotion Association Lexicon (NRC), auch EmoLex genannt, ist ein im Original englischsprachiges 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 Annotationen 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 assoziiert sein können. Die acht Emotionskategorien basieren auf der Emoti-

onstheorie 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 zeigen, dass durch die Übersetzung doppelte Wörter entstehen. Insgesamt hat das deutsche 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 assoziiert 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 enthalten, die für alle Kategorien den Wert 0 angegeben haben, also vollkommen neutral sind. Ein Beispiel ist das Wort „Avocado“. Abzüglich einiger weiterer Sonderfälle 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 „Bücher-

wurm“ 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, das 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; 2013) 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 assoziiert sind. 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.

2.3.4 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 von der 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:

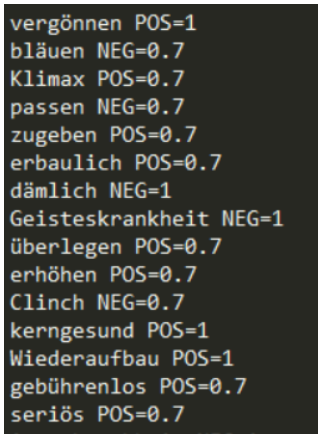


Abbildung 3: Ausschnitt CD

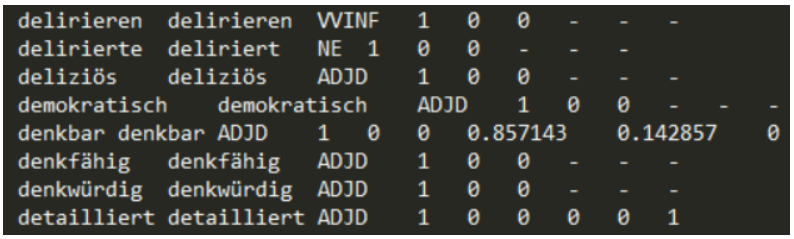
Das Lexikon enthält Wörter aus den Wortarten Nomen, Adjektiven und Verben sowie einige Phrasen. Insgesamt sind 2912 Einträge positiv annotiert, 4889 negativ und 594 neutral. Ferner sind noch 7 sogenannte Valenz-Shifter enthalten. Es handelt sich dabei um Wörter, die die Polaritätsausrichtung und -stärke 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 derartige Wörter. Ferner enthält es auch 124 Phrasen, also Einträge, die aus mehr als einem Wort bestehen.

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 2.2 angesprochen. Synsets sind semantische Konzepte zu denen mehrere Terme gehören können, die ein und dasselbe Konzept repräsentieren. Etwa 8000 Synsets wurden manuell bezüglich Polarität (positiv, negativ und neutral) sowie bezüglich Polaritätsstärke (ge-

ring, 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 wurden analog zu anderen Verfahren mit Seed-Words über einen Satz fester Polaritäts-Wörter weitere SBWs in einem Korpus akquiriert, die noch von menschlichen Annotatoren überprüft und bewertet wurden.

2.3.5 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. Artikel. Es ist nicht sicher, ob alle Angaben bezüglich Grundform und POS-Tag vollständig korrekt sind. Daraufhin folgen die Angaben zur Polaritätsklasse (positiv, negativ, neutral) und die eben erwähnten Wahrscheinlichkeiten. Folgender Screenshot illustriert den Aufbau:



```
delirieren  delirieren  VVINF   1   0   0   -   -   -
delirierte  deliriert   NE  1   0   0   -   -   -
deliziös    deliziös    ADJD    1   0   0   -   -   -
demokratisch    demokratisch    ADJD    1   0   0   -   -   -
denkbar denkbar ADJD    1   0   0   0.857143    0.142857    0
denkfähig   denkfähig   ADJD    1   0   0   -   -   -
denkwürdig  denkwürdig  ADJD    1   0   0   -   -   -
detailliert detailliert ADJD    1   0   0   0   0   1
```

Abbildung 4: Ausschnitt GPC

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. 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 evaluiert (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 wurde 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 deutschsprachigen 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.

2.3.6 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 Gründen 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 Ebene 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) empfeh-

len 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 Lexika SubjectivityClues (Wiebe et al., 2005) und SentiSpin (Takamura et al., 2005) erstellt. 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 2.4 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; Hölzer, Scheytt & Kächele, 1992) stammt aus dem Kontext der Psychologie und Depressionsforschung und umfasst etwa 25 000 flektierte Wörter mit Angaben für 8 Emotionskategorien, darunter 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 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).

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. Die Frage nach dem Nutzen des Einsatzes von Lexikon-basierten SA-Verfahren in der Dramenanalyse kann über die Auswahl ausreichend und detailliert beantwortet werden.

2.4 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 benennen 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 2.1) 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 (Tsytsarau & Palpanas, 2012; Ravi & Ravi, 2015). In Kapitel 2.1 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 (2005a) 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 (2005a) weisen selber darauf hin, dass diese Ergebnisse deutlich schlechter sind als emotionale Annotationen in anderen Bereichen. 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 Übereinstimmungen 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 wurden dabei als endgül-

tiger 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 (2005a) 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 positi-

ven, 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 (2005a) 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, 2005a). 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 (2005a) zeigt sich, dass die Mehrzahl des Textes als neutral bzw. nicht annotiert wurde (etwa 65-75%). Entgegen zu Alm und Sproat (2005a) 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 und eine fast perfekte unter Betrachtung zusammengehöriger Emotionen in Emotionsclustern fest. 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 (2005a), 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 (2005a) 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 (2005a) können derartiges Verhalten auch identifizieren und sehen das als Zeichen des breiten individuellen In-

terpretationsspielraums 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, das 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 2.3.3 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 Furcht, 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 Un-

terschieden 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 analysiert er die Emotionsverteilung eines großen Google-Book-Korpus entlang der Jahresachse und für verschiedene Nationen. So erkennt er zum Beispiel einen Anstieg von Furcht-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) beschäftigt 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 ein 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 größeren Verarbeitungsschritt auf lexikalischer oder textueller Ebene verwendet. Es werden die absoluten Valenzwerte durch einfache 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 2.3.3). 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) basieren auf 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 Mehr-

heitsentscheidung, 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.

3 Forschungsfrage 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 Interpreta-

tionen 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, das 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 For-

schungsagenda 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 (Mohammad, 2011; Nalisnick & Baird, 2013). 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 6 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 annotiertes Test-Korpus (Gold Standard; Turney, 2002; Wilson et al., 2005; Kaji & Kitsuregawa, 2007; Takala et al., 2014; Ribeiro et al., 2016; Maynard & Bontcheva, 2016). 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, 2005a). 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 grob zusammen. Es werden ferner die Kapitel angegeben, die sich mit der konkreten Ausarbeitung und Lösung der Arbeitsschritte befassen. In den jeweiligen Kapiteln werden die einzelnen Schritte im Hinblick ihres Nutzens und im Kontext der For-

schung noch mal explizit diskutiert. Des Weiteren werden, falls vorhanden, die jeweiligen produzierten Endprodukte der einzelnen Schritte angegeben.

Tabelle 1: Forschungsagenda

Arbeitsschritt	End- oder Zwischenprodukt	Kapitel
Analyse der Literatur und Forschung		Kapitel 2
Identifikation von einsetzbaren SA-Methoden		Kapitel 2 und Kapitel 5
Implementierung von SA-Methoden	Python-Programm-Dateien im Back-End	Kapitel 5
Implementierung eines Frameworks zur Vokabular-basierten Evaluation	Python Programm-Dateien im Back-End	Kapitel 6
Durchführung einer Vokabular-basierten Evaluation	Evaluationsergebnisse	Kapitel 6
Erstellung eines Gold-Standard-Korpus	Annotiertes Gold-Standard-Korpus	Kapitel 7
Analyse des Annotationsverhaltens	Statistische Auswertungen	Kapitel 7
Implementierung eines Frameworks zur SA-Evaluation	Python Programm-Dateien im Back-End	Kapitel 8
Durchführung einer SA-Evaluation	Evaluationsergebnisse → bestes SA-Verfahren	Kapitel 8
Entwicklung und Konzeption eines Front-Ends	Webanwendung	Kapitel 9
Analyse von Fallbeispielen		Kapitel 10

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 zahlreiche 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 deutschsprachigen Bereich ist kein Projekt bekannt, das 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.

4 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. Eine 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 Arbeit wurde der Parser jedoch noch erweitert und für den SA-Einsatz angepasst sowie kleinere Fehler behoben (siehe auch Kapitel 5.2). Folgender Screenshot illustriert für einen Teil die Art und Weise der XML-Auszeichnung in den TextGrid-Dramen:

```
<sp>
   <speaker xml:id="tg348.2.9.part1">JUNGFER OHLDIN.</speaker>
   <p xml:id="tg348.2.9.part2"> Sie können gewiß glauben, daß es mein Betrieb gar nicht
   gewesen ist. Die Heiraten werden im Himmel gestiftet, und wer wollte so gottlos sein,
   sich hier zu widersetzen?</p>
</sp>
<sp>
   <speaker xml:id="tg348.2.10.part1">CLITANDER.</speaker>
   <p xml:id="tg348.2.10.part2"> Da haben Sie recht. Die ganze Stadt lacht zwar über Sie;
   aber das ist das Schicksal der Frommen. Kehren Sie sich nicht daran. Ein Mann ist doch
   ein ganz nützlicher Hausrat.</p>
</sp>
<sp>
   <speaker xml:id="tg348.2.11.part1">JUNGFER OHLDIN.</speaker>
   <p xml:id="tg348.2.11.part2"> Ich weiß nicht, worüber die Stadt lachen sollte. Ist denn
   eine Heirat so was Lächerliches? die gottlose böse Stadt!</p>
</sp>
```

Abbildung 5: Ausschnitt TextGrid-Drama

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 wurde 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 oft 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, das 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 des 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.

Tabelle 2: Korpus – Dramen-Statistiken

Titel	Jahr	Gattung	Zahl der Repliken	Längste Replik	Replikenlänge (Avg)	Anzahl an Wörtern
Damon oder die wahre Freundschaft	1747	Lustspiel	183	274	40.35	7385
Der Freigeist	1749	Lustspiel	893	306	22.77	20338
Der junge Gelehrte	1748	Lustspiel	1038	411	23.96	24876
Der Misogyn	1748	Lustspiel	477	241	25.93	12369
Der Schatz	1750	Lustspiel	612	220	18.94	11596
Die alte Jungfer	1748	Lustspiel	470	117	19.71	9266
Die Juden	1749	Lustspiel	380	200	25.67	9755
Emilia Galotti	1772	Bürgerliches Trauerspiel	835	234	23.16	19342
Minna von Barnhelm oder Das Soldatenglück	1767	Lustspiel	1134	288	20.75	23538
Miss Sara Sampson	1755	Bürgerliches Trauerspiel	690	382	36.80	25395
Nathan der Weise	1779	Dramatisches Gedicht	1331	428	20.94	27884
Philotas	1759	Trauerspiel	181	775	37.93	6867

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

Tabelle 3: Korpus – Gesamt-Statistiken

Zahl der Repliken	Längste Replik	Replikenlänge (Avg)	Replikenlänge (Median)	Anzahl an Wörtern
8224	775	24.15	13.0	198611

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.

5 Back End: Sentiment Analysis

Der Großteil des Projekts besteht aus der Implementierung einer SA-Berechnungskomponente als Erweiterung des bestehenden Katharsis-Back-Ends. In den folgenden Kapiteln wird die grundsätzlich Idee und Konzeption geschildert und die einzelnen Programm-Bestandteile sowie alle möglichen Optionen knapp zusammengefasst. Für ausführliche Einsichten wird jedoch stets auf den Programm-Code im Back-End verwiesen.

5.1 Konzeption

Die grundsätzliche Konzeption der SA-Komponente basiert auf der Idee eine adaptierbare SA-Pipeline aufzubauen, deren einzelne Optionen, wie z.B. die Wahl des Lexikons, des Lemmatisierers manuell angepasst werden können. Zahlreiche Teilkomponenten mussten dafür entwickelt werden. Analog zum bisherigen Projekt, wurde die gesamte Back-End-Komponente mit der Programmiersprache Python entwickelt. Es werden jetzt grob die einzelnen Bestandteile geschildert, die dann ausführlicher in separaten Kapiteln erläutert werden. In den nachfolgenden Kapiteln wird dabei auch der Bezug zur Forschung hergestellt, die Einfluss auf die Konzeptions- und Entwicklungsentscheidungen hatte. Da die SA-Komponenten auf einige Programme des bisherigen Back-Ends angewiesen sind, ist eine komplette Trennung nicht möglich. Auch im Anhang wird das Gesamtprojekt weitergereicht. Es werden jedoch in den nächsten Kapiteln explizit die Dateien genannt, die erweitert wurden sowie die Dateien, die komplett

neu programmiert wurden, um die Abgrenzung zum bisherigen Katharsis-Projekt zu verdeutlichen.

Vor der grundsätzlichen Berechnung von SA-Metriken werden sowohl Verarbeitungsschritte für den zu analysierenden Text, also die Dramen, als auch die Lexika eingebaut. Diese werden in Kapitel 5.2 – 5.4 geschildert. Das Dramen-Modell von der Originalversion von Katharsis wurde angepasst und erweitert. Zahlreiche zusätzliche Attribute müssen in einem Pre-Processing-Schritt erstellt werden (Kapitel 5.2.1). Ein zentraler Teil ist dabei die Lemmatisierung des Dramentextes und die Abspeicherung der Lemmas (Kapitel 5.2.2). Auf der Lexikon-Seite werden die Lexikas aus ihrem Rohzustand in eine nutzbare Form übertragen, unnötige Daten gefiltert, und verarbeitbare Ausgabedaten erzeugt (Kapitel 5.3.2). Auch hier werden für den späteren Abgleich bei der SA Lemmatisierungsoptionen und -ausgaben eingebaut (Kapitel 5.2.2). Ferner wurde eine kombinierte Lexikonversion zur vereinfachten Berechnung aller Lexikon-Metriken und zur Berechnung kombinierter Werte erstellt (Kapitel 5.3.2). Als Optionen zur Optimierung der SA in Vorverarbeitungsschritten wurde die Möglichkeit einer Lexikonerweiterung mit historischen und linguistischen Wortvarianten (siehe Kapitel 5.3.3) sowie die Verwendung von verschiedenen Stoppwortlisten eingebaut (siehe Kapitel 5.2.3). Ferner kann zwischen verschiedenen Lemmatisierungsarten unterschieden werden (siehe Kapitel 5.3.4). In Kapitel 5.5 wird dann die Modellierung, Kalkulation und Ausgabe der Sentiment Analysis als beschrieben, sowie alle Metriken, die produziert werden.

5.2 Vorverarbeitung der Dramen

Vor der Durchführung der SA im Back-End müssen die vorhandenen Dramen vorverarbeitet werden um beispielsweise strukturelle, linguistische oder SA-spezifische Informationen anzufügen.

5.2.1 Allgemeine Vorverarbeitung

Die allgemeine Vorverarbeitung der Dramen beginnt bei den Dramen selbst. So mussten einige Dramen des Korpus geringfügig angepasst werden um mit der bisherigen Dramenmodell-Erstellung zu funktionieren. So wurde für Dramen die nur aus Szenen bestehen händisch eine Akt-Auszeichnung angefügt. Im Verlauf des Projekts wurden auch noch andere Probleme in der Annotation (z.B. fehlende Figuren) identifiziert und

ausgebessert. Die Roh-Dramen liegen im Ordner Lessing-Dramen-roh, die angepassten im Ordner Lessing-Dramen.

Das bisherige Tool parst die XML-Dramen und führt sie in ein spezielles Dramen-Modell als Pyhton-Objekt über. Das Dramen-Modell speichert Metadaten, strukturelle Informationen und Replikenstatistiken in einem eingebetteten Format. Das heißt auf höchster Ebene werden Dramen-Informationen gesichert, und das Drama besteht wieder aus Akt-Objekten die aus Konfigurations-Objekten bestehen, die wiederum aus Repliken-Objekten bestehen. Dieses Dramen-Modell als auch der Parser mussten angepasst und teilweise neu programmiert werden. Das Dramen-Modell und alle dazugehörigen Klassen in drama_models.py insofern, dass notwendige Sentiment-Attribute gemäß der SA-Modellierung hinzugefügt wurden. Aus diesem Grund wird im Folgenden stets zwischen dem herkömmlichen Dramen-Modell und dem erweiterten Dramen-Modell unterschieden. Das erweiterte Dramen-Modell beinhaltet die genannten SA-Informationen. In der Datei des Parsers drama_parser.py wurden Anpassungen zur korrekten Funktionalität mit dem gewählten Dramen-Korpus vorgenommen.

Das zentrale Vorverarbeitungsprogramm, das komplett neu geschrieben wurde ist nun aber sa_pre_processing.py. Die Klasse Drama_Pre_Processing implementiert dabei alle notwendigen Methoden und Verarbeitungsschritte zur Vorverarbeitung von einzelnen oder mehreren Dramen. Dafür muss den Methoden meist der Pfad zur XML-Datei übergeben werden. Es wird dann zunächst das oben beschrieben Dramen-Model mit dem bisherigen Parser erstellt und dann die Weiterverarbeitung durchgeführt und das erweiterte Dramen-Model zurückgegeben. Zentrale Methode hierzu ist preProcessAndLemmatize. Es werden in der Vorverarbeitung die Position der Repliken und Konfigurationen im Gesamtdrama aber auch allen anderen strukturellen Einheiten (z.B. im Akt) gesichert. Jede Replik wird insofern erweitert, dass der vorherige Sprecher gespeichert wird. Dies ist notwendig zur späteren Kalkulation von Sentiment-Beziehungen. Des Weiteren werden neue Sprecher-Objekte angefügt. Sprecher als Objekte mit dazugehörigen Sprecher-spezifischen Repliken werden bislang nur auf Dramen-Ebene gespeichert. Für die spätere SA werden Sprecher-Objekte pro Akt und Szene erzeugt und gesichert. Ferner wird noch, die durch Lemmatisierer berechnete Länge in Worten für jede strukturelle Einheit (Drama, Akt, Szene, Replik, Sprecher pro Drama, pro Akt, pro Szene) berechnet. Als letzter Schritt wird noch eine Lemmatisierung

aller Dramen durchgeführt und die dadurch gewonnen linguistischen Informationen (Part of Speech, Token, Lemma) für jede strukturelle Einheit an das Modell angefügt. Die erstellten Dramen können als Python-Dumps über die Pickle-Library als Pickle-Dateien ausgegeben und gespeichert werden. Pickle-Dateien sind Python-Objekte die über die Pickle-Library von Python einfach in diesem speziellen Python-Format abgespeichert und wieder abgerufen werden können. Diese Form der Abspeicherung wird im vorliegenden Projekt häufig für zeitintensive Algorithmen durchgeführt, die Dateien produzieren die nicht explizit eingesehen werden müssen. Ansonsten werden txt-Dateien oder ähnliches produziert. Die Generierung der Pickle-Datei ist sinnvoll, da die Lemmatisierung eine gewisse Zeit benötigt und so die Dramen im vorverarbeiteten Zustand zur Verfügung stehen. Das konkrete Vorgehen der Lemmatisierung wird im nächsten Kapitel geschildert. Die generierten erweiterten Dramen werden im Ordner Python/Dumps/ProcessedDramas gesichert und können von dort aus allen Python-Programmen einfach eingelesen werden. Je nach genutztem Lemmatisierer gibt es einen Ordner für mit textblob und treetagger verarbeitete Dramen.

5.2.2 Sprachverarbeitung – Lemmatisierung

Die Lemmatisierung von Text in Form von Dramentexten aber auch auf Seiten der Lexika ist möglicher Bestandteil der SA-Pipeline wird aber auch für vereinzelte andere Programme genutzt. Im Kontext dieses Oberkapitels wird die Lemmatisierungs zur Gewinnung von Lemmas der Dramen eingesetzt.

In diesem Kapitel wird die grundsätzliche Implementierung und Funktionalität der umgesetzten Lemmatisierung grob geschildert. Diese wird dann in kommenden Kapiteln bei denen die Lemmatisierung eingesetzt wird, nicht mehr ausführlich beschrieben sondern an diese Stelle verwiesen.

5.2.2.1 Idee

Lemmatisierung ist ein häufig eingesetzter Schritt zur Verbesserung der Leistung bei der Sentiment Analyse (Kennedy & Inkpen, 2006; Taboado et al., 2011; Peleja, Santos & Magalhaes, 2014; Ashgar et al., 2014; Hogenboom et al., 2015). Unter Lemmatisierung versteht man den Vorgang jedem Wort eines laufenden Textes seine Grundform zuzuweisen. Die Grundform eines Wortes wird im Weiteren auch Lemma genannt. Über den Prozess der Lemmatisierung wird die Flektionsform eines Wortes lexikographisch

auf seine Grundform reduziert (Lemma (Lexikographie), o. J.). Lemmatisierung ist ein wichtiger Prozess für die maschinelle Sprachverarbeitung und Anwendungsgebiete wie Information Retrieval, Wissensextraktion oder semantischer Analyse (Eger, Gleim & Mehler, 2016). Häufig wird Lemmatisierung zusammen mit Part-of-Speech-Tagging durchgeführt, also der Bestimmung der Wortart eines bestimmten Wortes (Schmid, 1995). Die POS wird im vorliegenden Projekt zwar erzeugt, gespeichert und häufig ausgegeben jedoch nicht direkt verwendet, kann aber in zukünftigen Projekten genutzt werden. In manchen SA-Projekten wird die POS genutzt (Denecke, 2008; Pak & Paroubek, 2010; Kouloumpis et al., 2011; Taboada et al., 2011; Cambria et al., 2013).

Speziell für das vorliegende Projekt kann die Lemmatisierung gewinnbringend sein, da vereinzelte Lexika nur Grundformen von Wörtern und keine flektierten Formen enthalten. Für die spätere Sentiment Analysis ist es jedoch wichtig zu jedem Wort eines Lexikons das passende Wort im Fließtext unabhängig von der Flektion zu finden. Es kann also über Lemmatisierung der wortbasierte Abgleich zwischen Lexikon und Dramentext verbessert werden, weswegen das Verfahren der Lemmatisierung vor allem bei lexikon-basierter SA Einsatz findet. Im vorliegenden Projekt werden zwei verschiedene Lemmatisierer eingesetzt. Einmal über die python-Library texblob der pattern-Lemmatisierer (De Smedt & Daelemans 2012). Bei textblob[1] handelt es sich um eine Python-Library zur Sprachverarbeitung speziell deutscher Spracher mit verschiedenen Klassen und Funktionen zur linguistischen Analyse von Text. Die Lemmatisierungskomponenten werden über die deutschsprachige Komponente des pattern-Lemmatisierers implementiert. Zum anderen wurde der Lemmatisierer und POS-Tagger treetagger von Schmid (1995) in Form eines Python-Wrappers eingesetzt[2]. Beide Lemmatisierer können leicht in ein Python-Programm integriert werden. Weitere Lemmatisierer für das Deutsche, die ohne größere Bearbeitungsschritte in Python verwendet werden können, konnten nicht identifiziert werden. Beide Lemmatisierer werden erfolgreich in der Forschung eingesetzt, der treetagger z.B. bei Monz und De Rijke (2001) im Information Retrieval und der pattern-Lemmatisierer bei De Fortuny et al. (2012) im Bereich des Text-Mining. Man hat sich dazu entschieden beide zu testen, um etwaige Probleme eines einzelnen Lemmatisierers in der späteren SA durch die Ver-

[1] https://pypi.python.org/pypi/textblob-de
[2] http://treetaggerwrapper.readthedocs.io/en/latest/

wendung des anderen auszugleichen. Ferner kann der grundsätzliche Nutzen der Lemmatisierung durch Evaluation zweier Lemmatisierer differenzierter analysiert und diskutiert werden.

Kritisch bezüglich der Lemmatisierung sei zu erwähnen, dass diese im Deutschen, das eine flektionsreiche Sprache darstellt, problematischer als in anderen Sprachen ist (Eger et al., 2016). Fehler in der Lemmatisierung sowohl auf Dramen-Text als auch auf Lexikon-Ebene können zu Problemen bei der SA führen. Als Beispiel sei die Reduktion verschiedener Wörter mit unterschiedlichen Sentiment-Bewertungen auf dieselbe Grundform genannt, die zur Kreation Lemma-spezifischer Ambiguitäten führt. Ein weiteres Problem ist, dass die Lemmatisierung im Fließtext häufig anders verläuft und zu anderen Ergebnissen führt als die Lemmatisierung auf Lexikonebene. Dies kann zu Problemen beim Wortabgleich der SA führen, die als Option Lemmatisierung auf beiden Ebenen durchführt. Einige SA-Projekte setzen Lemmatisierung nicht ein und nutzen vereinfachend die unbearbeiteten Wörter (Mohammad, 2011; Nalisnick & Baird, 2013). Auch diese Option wurde im vorliegenden Projekt umgesetzt und evaluiert.

5.2.2.2 Entwicklung

Die Lemmatisierung wird zentral in den Programm-Dateien lp_language_processor.py, lp_textblob.py und lp_treetagger.py. Die lp_language_processor.py gibt lediglich eine übergeordnete Steuer-Klasse Language_Processor vor. Objekte dieser Klasse werden in allen Projekten erzeugt, die die Sprachverarbeitung nutzen. Man übergibt dieser Klasse lediglich als String welchen Lemmatisierer man nutzen will. Die konkrete Sprachverarbeitung ist dann in den anderen beiden Dateien über die Klasse Text_Blob und Tree_Tagger umgesetzt. Obschon sich beide Klassen in der Struktur und einigen Methoden ähneln wurden zwei verschieden Klassen konstruiert, da beide integrierten Libraries sehr unterschiedlich funktionieren. Die finalen Datenstrukturen sind dabei jedoch gleich. Beide Klassen enthalten Attribute und Methoden zur Generierung von linguistischen Informationen. Als Attribute werden Tokens (also die unbearbeiteten im Text gefundenen Wörter), POS und Lemmas gesichert und zugreifbar gemacht. Zentrale Datenstruktur ist die Liste _lemmasWithLanguageInfo. Es handelt sich dabei um eine geordnete Liste jedes Wortes eines übergebenen Textes, dass die Worte als Tupel aus Lemma und Sprachinformation gespeichert ist. Als Sprachinformation wird wiederum ein Tupel aus dazuge-

hörigen Token und POS verstanden. Sowohl pattern als auch treetagger könne auch POS erstellen. Über die Methode processText kann ein Text übergeben werden und obige Datenstrukturen werden erstellt und dem Klassen-Objekt von außen zugreifbar hinzugefügt. Die Erstellung der Strukturen ist dabei bei den einzelnen Lemmatisierern unterschiedlich und wird über verschieden Methoden und Verarbeitungsschritte umgesetzt. Sie wird hier nicht weiter vertieft, kann aber im Programm-Code eingesehen werden. Es findet dabei aber unter anderem auch eine Filterung einiger Sonderzeichen im Text vor, die zu Problemen bei der Lemmatisierung geführt haben. Die Methode processTextTokens erstellt nur Token-Listen und wird für einige Programme gebraucht. Die Methode processTextFully erstellt komplexere Speicherstrukturen (Dictionaries), die für Programme der vokabularbasierten Evaluation notwendig sind (siehe Kapitel 6). Ferner wird auch die Stoppwort-Verwendung und Generierung in den Lemmatisierungs-Klassen verwaltet (siehe Kapitel 5.2.3).

5.2.3 Stoppwortlisten

Neben der Lemmatisierung ist auch der Einsatz von Stoppwortlisten auf Dramentext-Ebene eine Optimierungsoption. Der Einsatz von Stoppwortlisten kann dabei als sowohl bezogen auf die Dramenebene, als auch bezogen auf die Lexikon-Ebene als Option betrachtet werden. Im Programm-Code wird der Einsatz der Stoppwortlisten tatsächlich so implementiert, dass im Text gefundene Stoppwörter bei der SA ignoriert werden und die Lexika also nicht explizit nach Stoppwörtern gefiltert werden. Aus diesem Grund wird dieses Verfahren im jetzigen auf die Dramenebene bezogenen Kapitel erläutert.

5.2.3.1 Idee

Unter Stoppwörtern versteht man Wörter, die sehr häufig auftreten und gewöhnlich keine Relevanz für die Erfassung des Dokumentinhaltes haben. Es handelt sich bei Stoppwörtern meist um die am häufigsten in einer Sprache vorkommenden Wörter. Im Deutschen sind viele Stoppwörter Artikel, Konjunktionen oder Präpositionen (Stoppwort, o. J.). Eine Stoppwortliste ist eine Sammlung derartiger Wörter. In der SA werden Stoppwortlisten als Vorverarbeitungsschritt für den Text verwendet (Hu und Liu, 2010, Asghar et al., 2014) um beispielsweise die Zahl an notwendigen Kalkulationen zu reduzieren. Saif et al. (2014) nutzen Stoppwortlisten aber tatsächlich nun zur direkten

Optimierung der SA. So können sie feststellen, dass der Einsatz dieser durch die Reduktion der genutzten Features ihres ML-Algorithmus, tatsächlich die Genauigkeit der SA steigert. Sie experimentieren dabei erfolgreich mit verschiedenen Arten Stoppwortlisten.

Diese grundsätzliche Idee der Nutzung von Stoppwortlisten zur SA-Optimierung von Saif et al. (2014) wird im vorliegenden Projekt aufgegriffen und für den Einsatz bei Lexikon-basierten Verfahren begründet. Die Kalkulation für die hier genutzten SA-Verfahren basiert, vereinfacht betrachtet, stets auf der Summierung von Sentiment-konnotierten Wörtern. Befinden sich nun Stoppwörter in den Lexika fließt eine übermäßige Menge an SBWs in die Kalkulation ein, die das Endergebnis verfälscht. Dies ist insbesondere der Fall wenn das Stoppwort fälschlicherweise als Sentiment-tragend ausgezeichnet ist. Andere Probleme können Wörter verursachen, die aufgrund der Sprachwahl von Lessing besonders häufig verwendet werden, die tatsächlich keine konkreten Stoppwörter sind und ohne Kontext als korrekt Sentiment-tragend betrachtet werden können. Auf Basis der häufigen inflationären Verwendung solcher Wörter kann dennoch ein problematischer Einfluss auf die Endkalkulation entstehen. In der Tat konnten im Projekt für die genannten Fälle konkrete Beispiele gefunden werden, die je nach Annotation des Wortes zu einer übermäßig positiven oder übermäßig negativen Bewertung von Repliken führen.

Aus den genannten Gründen wurde also die Verwendung von Stoppwortlisten implementiert und auf ihre Leistung hin evaluiert. Es wurden nach ähnlichem Schema wie bei Saif et al. (2014) verschieden Stoppwortlisten erstellt. Als Standard-Stoppwortliste wurde eine deutschsprachige Stoppwortliste verwendet, die mit dem freien Information-Retrieval-System Solr mitgeliefert wird[3]. Es handelt sich um eine verhältnismäßig kleine Stoppwortliste von lediglich 231 Einträgen, die die eindeutigsten Stoppwörter der deutschen Sprache enthält. Die Liste wurde noch mit den großgeschriebenen Wörtern der Stoppwörter erweitert, um bei fehlender Lemmatisierung Stoppwörter am Satzanfang abzufangen. Die finale Liste hat also 462 Einträge und wird im Folgenden standardList genannt. In der Tat konnte konstatiert werden, dass einige Lexika Wörter dieser Liste als SBWs enthalten z.B. das GPC die Wörter „dem“, „für“ oder „wer“ oder das BAWL-R Wörter wie „machen“ und „können“. Über die

[3] http://lucene.apache.org/solr/

Filterung des Dramentextes mit der standardList wird der Einfluss der Verwendung dieser Wörter untersucht.

Einer Idee von Saif et al. (2014) folgend wurde die standardList mit sonstigen sehr häufigen Wörtern des Gesamtkorpus erweitert. Dazu wurden mit den in Kapitel 6 beschriebenen Programmen die 100 häufigsten Wörter des Gesamtkorpus, die nicht ohnehin bereits Stoppwörter sind, akquiriert und die standardList damit erweitert. Es handelt sich also um Wörter, die übermäßig häufig vorkommen und aus diesem Grunde einen verzerrenden Einfluss auf die Kalkulation haben können. Diese erweiterte Liste wird als enhancedList bezeichnet. Es handelt sich um 562 Einträge. Groß- und Kleinschreibung wurde dabei nicht beachtet, da die Groß- und Kleinschreibung durch die Häufigkeit des Wortes vorgegeben wird, d.h. zum Beispiel, dass ein klein geschriebenes sehr häufiges Wort kommt nicht ausreichend häufig großgeschrieben vor, da es sonst bereits in der Liste der häufigsten 100 Wörter enthalten wäre. In der Tat konnten auf diese Weise einige Wörter in den Lexika identifiziert werden, die zu Problemen führen können. Als Beispiele seien die Wörter „Vater“ und „Tochter“ im BAWL-R oder die Wörter „Fräulein“ und „Herr“ im NRC. Diese können zwar als Sentiment-tragend interpretiert werden, werden aber von Lessing derart häufig und inflationär in seiner Sprache verwendet, dass man einen Verlust der Sentiment-Bedeutung annehmen kann. Insbesondere das Wort Fräulein, dass aus heutiger Sicht als höfliche Anrede positiv konnotiert ist, wird von Lessing als herkömmliche Anrede der Zeit verwendet. Dieses Beispiel verdeutlicht auch die Probleme der Anwendung moderner SA-Lexika auf ältere Texte.

Neben der enhancedList wurde dann noch die sogenannte enhancedFilteredList erstellt. Diese wurde durch manuelle Filterung von Wörtern in der enhancedList erstellt. Es wurden nach eigenem Ermessen Wörter entfernt, die zum einen trotz übermäßiger Häufigkeit als eindeutig Sentiment-tragend betrachtet wurden. Als Beispiele seien die Wörter „Liebe“, „Glück“ und „Herz“ genannt. Zum anderen wurden Wörter gefiltert, die Sentiment-tragend sein können, aber fälschlicherweise durch die Nutzung von Großschreibung im ersten Schritt zur Erstellung der standardList hinzugefügt wurden. Ein Beispiel hierfür ist das Wort „Macht“ vom Stoppwort „macht“ oder auch das Wort „Würde“ vom Stoppwort „würde. Über die Erstellung dieser Liste kann eine

zusätzliche manuell angepasste Form einer Stoppwortliste auf ihre Leistung hin evaluiert werden.

Die SA wurde so implementiert, dass man alle drei Stoppwortlisten, aber auch die Verwendung keiner Stoppwortliste als Option auswählen kann. Auf diese Weise kann der Nutzen von Stoppwortlisten in Interaktion mit den einzelnen Lexika und anderen Verfahren differenziert untersucht werden. Alle Stoppwortlisten mit zusätzlichen Informationen zur Erstellung findet man im Anhang im Ordner Stopwords.

5.2.3.2 Entwicklung

Alle programmiertechnischen Bestandteile zur Nutzung der Stoppwortlisten wurde in den Dateien lp_textblob und lp_treetagger implementiert. Über die Methode initStopwords kann der Name der Stoppwortliste übergeben werden und die jeweilige Stoppwortliste wird initialisiert. Es wird auch eine lemmatisierte Form der Stoppwortliste erstellt für den Abgleich von lemmatisierten Text mit den Stoppwörtern. Die lemmatisierte Stoppwortliste enthält die entstandenen Lemmas und zusätzlich die ursprünglichen Tokens um mögliche Lemmatisierungsfehler durch die isolierte Lemmatisierung einzelner Wörter abzufangen und die ursprünglichen Stoppwörter in jedem Fall zu filtern. Die Stoppwörter werden in den Attributen _stopwords und _stopwords_lemmatized gespeichert. Auf diese kann von außen zugegriffen werden. Über die Methoden removeStopwordsFromTokens und removeStopwordsFromLemmas können die Stoppwörter der übertragenen Texte aus den linguistischen Datenstrukturen entfernt werden.

5.3 Lexika-Verarbeitung

Auf Lexikon-Ebene finden können verschieden Verarbeitungsschritte statt. In der grundsätzlichen Verarbeitung werden die einzelnen Lexika eingelesen, angepasst, in eine verarbeitbare Datenstruktur überführt und Ausgaben produziert. Es wurde auch eine kombinierte Lexikon-Version erstellt und genutzt. Als fortgeschrittene Methoden der Lexikon-Verarbeitung sind eine Erweiterung mit linguistischen und historischen Wortvarianten sowie eine bereits angesprochene Lemmatisierung möglich.

5.3.1 Grundsätzliche Verarbeitung und Lexikon-Auswahl

Die allgemeine Verarbeitung jedes einzelnen Lexikons wird in separaten Programm-Dateien durchgeführt. Zwar sind einige grundsätzliche Schritt ähnlich, doch die sehr

unterschiedlichen Formate und Annotationen für jedes Lexikon machten die Entwicklung von Methoden notwendig, die angepasst auf jedes Lexikon sind.

Die einzelnen Dateien zur Lexikonverarbeitung sind nach dem jeweiligen Lexikon benannt. Es handelt sich um die Dateien lexicon_bawl.py, lexicon_clematide_dictionary.py, lexicon_german_poalarity_clues.py, lexicon_nrc.py und lexicon_sentiWS.py. An dieser Stelle kann nicht im Detail auf die Implementierung aller Verarbeitungsketten eingegangen werden, dafür wird wieder auf den Programm-Code im Anhang verwiesen.

Jedes Programm bildet eine eigene Klasse, die die Rohdaten des Lexikons einliest und in eine in Python nutzbare Dictionary-Struktur umwandelt. Diese besteht aus Schlüssel-Wert-Paaren in Form der Wörter des Lexikons als Schlüssel und der Sentiment-Werte als Werte. Je nach Art des Lexikons werden die Sentiment-Werte als Polaritätsstärken oder wieder als Schlüssel-Wert-Paare mit den Sentiment-Kategorien als Schlüssel und den einzelnen Ausprägungen als Werte gespeichert. Das jeweilige Dictionary wird in dem Attribut _sentimentDict gesichert. Die lemmatisierte Form des Lexikons in dem Attribut _sentimentDictLemmas. Je nach Lexikon liegen verschiedene Zwischenschritte zum Beispiel zum Herausfiltern von Lexikoneinträgen vor. Als Beispiele sei das Entfernen von Sonderzeichen im GPC genannt, hier sind Sonderwörter, die aus Satzzeichen bestehen als Einträge enthalten oder das NRC, in dem komplett neutrale Wörter ohne jegliche Annotation enthalten sind. Neutrale Einträge werden generell entfernt; im vorliegenden Projekt wird zur Vereinfachung der SA-Aufgaben nicht mit der Klasse Neutral für die Polarität gearbeitet.

Die lemmatisierte Form jedes Lexikon-Eintrages wird durch Nutzung der Lemmatisierungsprogramme, die bereits in Kapitel 5.2.2 erörtert wurden, erlangt. Dabei ist jedoch stets zu beachten, dass die isolierte Lemmatisierung eines Wortes problematisch für den Lemmatisierer sein kann, da dieser häufig auf den umliegenden Text für eine korrekte Lemmatisierung benötigt. Dennoch kann mit der lemmatisierten Form zumindest untersucht werden ob, trotz der Fehlerhaftigkeit, eine Lemmatisierung auf Lexikon-Ebene beim Abgleich mit lemmatisierten Text gewinnbringend ist. Für manche Lexika sollte dies grundsätzlich nicht notwendig sein, da diese korrekte Wortformen enthalten.

In verschiedenen Verarbeitungsschritten kann es vorkommen, dass gleiche Wörter mit verschiedenen Sentiment-Angaben erzeugt werden. Dies kann bereits dadurch entstehen, dass gleiche Wörter in den Roh-Formen des Lexikons mit unterschiedlichen Annotationen vorliegen (Beispiele). Es kann jedoch auch im Schritt der Lemmatisierung geschehen, indem flektierte Formen eines Wortes, die verschieden Sentiments haben, auf die gleiche Grundform reduziert werden und dann gleich geschrieben sind aber unterschiedliche Sentiment-Ausprägungen haben. Dies kann insbesondere bei den Lexika vorkommen die flektierte Formen enthalten. Bei jeder Doppelung eines Wortes wird für jedes Lexikon über verschieden Methoden entschieden, welche Sentiment-Ausprägung für die finale Lexikon-Form gespeichert wird. Grundsätzlich wird bei Polaritätsstärken stets die stärkere Sentiment-Ausprägung gesichert, also bei zwei negativen Wörtern diejenige Ausprägung mit dem höchsten Ausschlag oder bei einem positiven und negativen Sentiment für zwei gleiche Wörter dasjenige, dessen absoluter Wert am höchsten ist. Die heuristische Idee dabei ist, dass das stärkere Sentiment eher Ausdruck für das am häufigsten auftretende Sentiment ist. Für den Fall des NRC wird diejenige Sentiment-Ausprägung mit den meisten Annotationen für Polaritäten und Emotionskategorien. Sollten die Polaritätsstärken gleich sein oder das Lexikon nur Angaben zu Polaritätsklassen enthalten, wird diejenige Polarität gewählt, die am seltensten bezogen auf die sonstigen Polaritäten im Korpus, vorkommt. Die gewählten Verfahren sind kritisch zu betrachten. Zur korrekten Funktionalität musste jedoch ein Umgang mit ambigen Angaben gewählt werden. Relativierend ist anzumerken, dass es sich jeweils nur um einen minimalen Prozentsatz, der im gesamten Lexikon enthaltenen Wörter handelt, die über verschieden Arbeitsschritte als ambig identifiziert werden.

Die Lexikon-Programme bieten auch noch Methoden zur Ausgabe der verarbeiteten und transformierten Lexika nach einem geordneten Schema. Es handelt sich um Listen mit per Tab getrennten Annotationsangaben für zeilenweise Lexikon-Einträge. Es wird die normal verarbeitbare Form als Tokens und die lemmatisierte Form mit Lemmas der Lexika ausgegeben. Die lemmatisierten Formen gibt es zweimal, nämlich für jeden Lemmatisierer, pattern und treetagger, eine Datei. Die jeweiligen Dateien befinden sich im Ordner SentimentAnalysis/TransformedLexicons im Anhang. Die lemmatisierten Dateien werden von den Klassen selbst zur beschleunigten Initialisie-

rung des lemmatisierten Lexikons genutzt. Ferner wird in den Programmen auch die DTA-Erweiterung und Ausgabe umgesetzt. Diese Funktionalität wird noch in Kapitel 5.3.3 besprochen.

Das Programm lexicon_handler.py ist nun zentrale Steuereinheit für die Verwendung von Lexika. Über die Klasse Lexicon_Handler können verschieden Lexika initialisiert werden ohne dass auf die Originalprogramme zugegriffen werden müssen. Dabei wird dem Lexicon_Handler der Name des Lexikons und des Lemmatisierers übergeben. Über die Attribute _sentimentDict und _sentimentDictLemmas kann man auf die oben beschriebenen Sentiment-Lexika zugreifen. Über das Programm wird auch die Lexikonkombination umgesetzt, die noch im nächsten Kapitel erläutert wird.

5.3.2 Lexika-Kombination

Als weiteres fortgeschrittenes Lexikon-basiertes Verfahren wurde ein kombiniertes Lexikon aus den bestehenden Lexika erzeugt und verschiedene dazugehörige Metriken definiert und berechnet. Das kombinierte Lexikon und seine Metriken werden nicht als explizite Option der SA betrachtet, sondern als Teil der Option der Lexikon- und Metrik-Auswahl.

5.3.2.1 Idee

Die grundsätzliche Idee mehrere SA-Lexika zu einem größeren Gesamt-Lexikon zu kombinieren wird von Emerson und Declerck (2014) aufgegriffen. Sie können erfolgreich mehrere Lexika zu einem verbinden und damit nachweisliche Verbesserungen in der SA im Vergleich zu den Einzel-Lexika erreichen. Der Vorteile eines kombinierten Lexikons ist die Ausdehnung des vorhandenen Wortschatzes. Nachteile sind der Informationsverlust beim Umgang mit Ambiguitäten und der Zusammenführung verschiedener Annotationssysteme für Sentiments. Emerson und Declerck (2014) verwenden komplexe mathematische Modelle und Normalisierungsverfahren, um mit den unterschiedlichen Annotationsformaten der einzelnen Lexika optimal umzugehen. Die grundsätzlichen Ansätze von Emerson und Declerck (2014) werden hier aufgegriffen, jedoch eine einfachere Umsetzung entwickelt.

Um die konkrete Idee zu erläutern, müssen Teile der konkreten Programmierung aufgegriffen werden. Das kombinierte Lexikon ist zunächst in den vorliegenden Programmen eine simple Zusammenführung der Angaben aller Lexika; es findet zunächst

keine Ersetzung oder Anpassung von Werten statt. Dieses simple kombinierte Lexikon, im folgenden CombinedLexicon, genannt enthält jedes Wort aller Lexika mit Sentiment-Angaben als Schlüssel-Wert-Paare, d.h. zu jedem Wort ist ein Wert gesichert, der wiederum ein Schlüssel-Wert-Paar mit dem Namen des Lexikons und den jeweiligen Polaritätsangaben enthält. Auf diese Weise werden für alle Wörter alle Sentiment-Angaben aller Lexika unverändert und strukturiert gesichert. CombinedLexicon ist also eine simple Dictionary-Datenstruktur zur Speicherung aller Wörter und Sentiment-Werte der verwendeten Lexika. Diese Struktur wird in zahlreichen Programmen vereinfachend auch zur Kalkulation einzelner Lexika-Metriken verwendet, da man so für alle Lexika, die jeweiligen Metriken auf einmal berechnen kann.

Der konkrete Mehrwert dieser Datenstruktur und die Kombination von Lexika finden erst auf der Kalkulationsebene von Metriken statt. Es beschränkt sich im Gegensatz zu Emerson und Declerk (2014) auf alle Lexika-spezifischen Term-Zähl-Metriken, also auf alle Metriken, die pro SBW lediglich angeben, ob es zur positiven oder negativen Klasse gehört. Diese Metrik wird für jedes Lexikon als dichotom bezeichnet, da sie jeder Polaritätsklasse einen Wert von 1 oder 0 zuweist, je nachdem ob ein SBW negativ (negativ: 1; positiv: 0) oder positiv ist (negativ: 0, positiv: 1). Metriken über Polaritätsgewichte werden nicht genutzt. Für jedes Lexika liegt eine derartige Annotation vor oder wird rechnerisch erzeugt. Es werden zwei kombinierte Metriken definiert, um die Lexikonkombination umzusetzen. Für die eine Metrik wird jede dichotome Polaritätsangabe des gesamten Lexika-Wortschatzes über die Datenstruktur CombinedLexicon zur Kalkulation genutzt. Ambiguitäten, also unterschiedliche Polaritätsangaben in verschiedenen Lexika werden zunächst nach Mehrheitsentscheid gelöst, d.h. es wird diejenige Polarität gewählt, die in der Mehrzahl der Lexika auftritt. Bei einem Unentschieden entscheidet das Lexikon, das gemäß der in Kapitel 8 durchgeführten Evaluation die bessere Erkennungsleistung zeigt. Bei einer zweiten Metrik werden nur diejenige Worte als SBWs, die in mindestens drei Lexika mit eindeutiger Polarität vorkommen. Letztgenannte Methodik reduziert den Wortschatz aufgrund des starken Kriteriums stark. Das CombinedLexicon liegt als auf den Token basierten Wörtern vor, wird aber auch nach gleichem beschriebenem Schema lemmatisiert erzeugt. Weitere ausführliche Erläuterungen zur Kalkulation der Metriken findet man in Kapitel 5.5.

Durch die Entwicklung des CombinedLexicon als simple Datenstruktur und die Definition der genannten Metriken soll der Nutzen und Einsatz von kombinierten Lexikon-Verfahren im vorliegenden Projekt untersucht werden.

5.3.2.2 Entwicklung

Die Genese des zusammengesetzten Lexikons CombinedLexicon findet in der Programm-Datei lexicon_handler.py in einem mehrstufigen Prozess statt. Zunächst werden über verschieden Methoden die Schlüssel, also die Lexikon-Wörter, für das spätere Dictionary zusammen gesammelt. Dies vereinfacht die spätere Erstellung des Lexikons. Die gesammelte Liste aller Lexika-Wörter wird als Python-Pickle-Datei im Ordner Dumps/LexiconKeys/ zwischengespeichert. Es wird dabei eine Schlüssel-Datei für die Token-Versionen der Lexika-Wörter, aber auch für die Lemmas (treetagger und pattern/textblob) erstellt und gesichert. Über die zentrale Methode combineSentiment-Lexica wird dann die kombinierte Datenstruktur CombinedLexicon erstellt, dazu wird nach den Schlüssel-Wörtern in den Einzellexika gesucht. Liegt ein Wort im Lexikon vor, wird ein Schlüssel-Wert-Paar aus dem Namen des Lexikons und den Sentiment-Ausprägungen erstellt. Dem Wortschlüssel wird der Lexikonname als Schlüssel zu den Ausprägungen gegeben. Auf diese Weise werden alle Wörter mit allen Lexikon-spezifischen Sentiment-Angaben gespeichert. Die finale Datenstruktur wird wieder als Pickle-Datei abgespeichert (Ordner: Dumps/CombinedLexicons). Es werden Lexika für Tokens und pro Lemmatisierer erstellt. Ferner werden auch einsehbare txt-Dateien im Ordner SentimentAnalysis/TransformedLexicons/ wie bei herkömmlichen Lexika kreiert und hinterlegt. Über die Übergabe von „CombinedLexicon“ als String wird dieses wie jedes andere Lexikon initialisiert und, wie üblich über die Attribute _sentimentDict und _sentimentDictLemmas die Schlüssel-Wert-Paare zugreifbar gemacht. Das CombinedLexicon kann nach dem ähnlichen Schema auch in Form einer DTA-Erweiterung in der Datei erzeugt werden (siehe auch nächstes Kapitel).

Die konkrete Berechnung aller Metriken für die tatsächliche Lexikon-Kombination findet wie für alle Metriken in der Programm-Datei sa_calculator statt und wird noch mal in Kapitel 5.5 angesprochen und aus programmier-technischer Sicht besprochen.

5.3.3 DTA – Erweiterung

Die DTA-Erweiterung ist im Folgenden die Erweiterung der einzelnen SA-Lexika mit linguistischen historischen Varianten über ein Tool des deutschen Textarchivs von Jurish (2012). Es handelt sich um eine Option für die SA-Durchführung. In den folgenden Abschnitten werden die grundsätzliche Idee und die technische Umsetzung erläutert.

5.3.3.1 Idee

Die Erweiterung und Erstellung eines SA-Lexikons ist eine Methodik aus der SA-Forschung. Meist wird das bestehende Set an Wörtern mit Synonymen oder anderen mit den Set-Wörtern gleichzeitig auftretenden Wörtern erweitert. Für die vorliegende Arbeit liegt die Vermutung nahe, dass der Wortschatz des 17. Jahrhunderts und die poetische Ausdrucksweise in einem Drama stark von den Wörtern der verwendeten Sentiment-Lexika abweicht, da diese auf Basis von modernen Online-Lexika (Vo et al., 2009) Produkt- (Remus et al., 2010) und New-Korpora (Clematide & Klenner, 2010) oder über Crowdsourcing erstellt wurden (Mohammad & Turney, 2010). Es wurden Unterschiede in der Orthographie aber auch in der Lemmatisierung erwartet. Aus diesem Grund wurde nach einer Lösung gesucht, historische linguistische Varianten in die Lexika zu integrieren. Ein vom Deutschen Text-Archiv online zugreifbares Web-Tool, das Jurish (2012) in seiner Dissertation entwickelt hat, stellt diese Funktionalität zur Verfügung. Dem Tool können über verschieden Dateiformate Wörter übergeben werden und es gibt die jeweiligen historischen linguistischen Varianten in einem strukturierten Format aus. Folgender Screenshot zeigt das Beispielwort „Gebrechlichkeit“ und die Rückgabe, die man vom Tool bei Übertragung dieser Query und bei korrekter Einstellung der Term Expansion erhält:

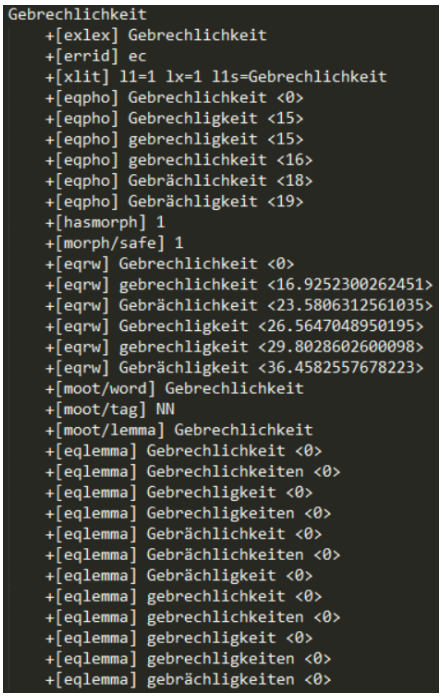


Abbildung 6: Beispiel DTA-Format

Relevant für die Lexikon-Erweiterung sind die historischen linguistischen Varianten des Wortes die über die Annotation [eqpho], [eqrw] und [eqlemma], die sonstigen Bestandteile können ignoriert werden. Unter [eqpho] werden phonetische Varianten, unter [eqrw] orthographische Varianten und unter [equlemma] mögliche flektierte Formen und historische Lemmaformen angegeben. Man erkennt deutlich, dass verschiedene Formen nützlich sein können bei der Erkennung von Wörtern in eine veralteten poetischen Text, da historische Schreibweisen des Wortes Gebrechlichkeit wie „Gebrächlichkeit“ nun abgefangen werden können. Als weiteres konkretes Beispiel sei das Wort „betriegen“ aus dem Korpus erwähnt, das Lessing sehr häufig verwendet und das mit dem in einigen Lexika enthaltenen negativ konnotierten betrügen zusammenhängt. Durch die Erweiterung mit dem Tool von Jurish (2012) können derartige Varianten nun erkannt werden. Ferner sei noch zu beachten, dass über die Lemma-Angabe des Tools eine explizite Form von Lemmatisierung für ein Wort stattfindet, da das Lexikon mit zusätzlichen Lemma-Varianten erweitert wird. Inwiefern dies die normale Lemmatisierung über pattern/textblob und treetagger überflüssig macht, wird noch bei der Evaluation in Kapitel 8 besprochen.

Für jedes Wort werden die Varianten bezogen und die Varianten mit den gleichen Sentiment-Ausprägungen gesichert. Man erkennt am obigen Beispiel deutlich, dass diese Form der Lexikon-Erweiterung das Lexikon sehr stark vergrößert. Als Beispiel

sei hier das BAWL-R genannt, das im Original 2842 Einträge besitzt und durch Lexikon-Erweiterung 75436 Einträge besitzt. Mit auftretenden gleichen Wörtern und Erweiterungen wird ähnlich verfahren wie mit gleichen Wörtern bei der Lexikon-Verarbeitung (siehe Kapitel 2.3).

Es wird im vorliegenden Projekt mittels Evaluationsverfahren untersucht, inwiefern eine derartige Lexikon-Erweiterung einen Mehrwert für die SA in der Dramenanalyse liefert.

5.3.3.2 Entwicklung

Die zentrale Programmierlogik für die DTA-Erweiterung befindet sich in der Programm-Datei lexicon_dta_enhancement. Bevor dieses Programm eingesetzt werden konnte, wurden die Wort-Erweiterungen über das Tool (2012) manuell über eine Übertragung aller Lexikon-Wörter als txt-Datei durchgeführt. Dies musste über vier Dateien unternommen werden, um das Datenlimit nicht zu überschrieben. Die erhaltenen Daten wurden mit den Wort-Erweiterungen abgesichert. Die Dateien befinden sich zur Einsicht im Ordner SentimentAnalysis/DTA-Output/FetchedDTAData.

Über die Klasse DTA_Handler wird im Attribut _wordSynonymsDict ein Dictionary bestehend aus Wörtern und ihren erhaltenen linguistischen historischen Varianten gespeichert. Dieses wird durch eine Methode, die die übertragenen DTA-Dateien einliest und verarbeitet, erstellt. Über die Methode extendSentimentDictDTA kann nun ein Lexikon explizit mit den Worten erweitert werden. Bei Wort-Doppelungen wird wie im vorigen Kapitel bereits angesprochen genauso verfahren wie bei der normalen Lexikon-Erweiterung. Die Methoden zur Auswahl der dann geltenden Sentiment-Ausprägung befinden sich deswegen auch in der Klasse. Alle Doppelungen mit Änderungen können im Ordner DTA-Output/AdditionalInfo inspiziert werden. In der Klasse können die erweiterten Lexika ferner auch als txt-Dateien ausgegeben werden. Die erweiterten Lexika inklusive die Erweiterung für alle Lexikon-Wörter als erweiterte Form von CombinedLexicon befinden sich im Anhang im Ordner Sentiment-Analysis/TranformedLexicon als strukturierte txt-Dateien.

Die jeweiligen lemmatisierten Versionen der einzelnen Lexika mit DTA-Erweiterungen werden in den Lexikon-Programmen (also z.B. lexicon_sentiWS.py usw.) für jeden Lemmatisierer erstellt und auch als einsehbare txt-Datei in obigem Ordner gesichert. Über das lexicon_handler.py-Programm werden die DTA-

Erweiterungen für das CombinedLexicon erstellt. Wenn man ein DTA-erweitertes Lexikon verwenden will, kann man dieses dem Lexikon-Handler als Lexikon-Namen übergeben. Die DTA-erweiterten Lexikon-Namen haben stets die Form LexikonNameDTAExtended, also zum Beispiel GPC-DTAExtended. Die einzelnen Lexika und die Lexicon_Handler-Klasse für CombinedLexicon können dann die DTA-Erweiterung über den DTA-Handler durchführen und die lemmatisierten Versionen aus Performanz-Gründen über die zuvor erstellten Dateien auch generieren. Wie üblich kann man die Lexika dann über die Attribute _sentimentDict und _sentimentDictLemmas beziehen.

5.3.4 Sprachverarbeitung – Lemmatisierung

Auch auf der Lexikon-Ebene wird die Option einer Lemmatisierung angeboten. Im Folgenden werden die verschiedenen Optionen der Ebenen der SA auf denen die Lemmatisierung stattfinden kann illustriert und begründet.

Für diesen Zweck werden dieselben Lemmatisierer und Programme, wie bereits in Kapitel 5.2.2 beschrieben, eingesetzt. Die konkrete Implementierung wird an dieser Stelle nicht erläutert.

5.3.4.1 Idee

Es wurde bereits angesprochen, dass Lemmatisierung häufig in SA-Pipelines eingesetzt wird. Bei Lexikon-basierten Verfahren bietet sie sich an, um den Abgleich des Lexikon-Wortes, das häufig in seiner Grundform vorliegt, mit einer flektierten Form im Text zu ermöglichen. Einige Lexika beinhalten zwar schon flektierte Formen (z.B. SentiWS und GPC), diese sind jedoch oft unvollständig oder unkorrekt, was beispielsweise im Fall des Lexikons SentiWS explizit mitgeteilt wird (Remus et al., 2010), weswegen auch dann eine Lemmatisierung noch immer gewinnbringend für den korrekten Wortabgleich sein kann. Man kann die Lemmatisierung jedoch auch kritisch betrachten.

Es wurden drei Möglichkeiten identifiziert und im vorliegenden Projekt implementiert um mit der Lemmatisierung umzugehen. Als erste Möglichkeit wird weder auf Text noch auf Lexikon-Ebene Lemmatisierung durchgeführt. Es werden also nur die rohen Token betrachtet. Man kann einige Projekte identifizieren, die in der SA selbst bei Lexikon-basierten Verfahren keine Lemmatisierung anwenden (Mohammad, 2011; Nalisnick & Baird, 2013). Auffällig ist der mögliche Informationsverlust der

durch Grundformreduktion entsteht. Dies konnte auch im vorliegenden Projekt bei der Verarbeitung der Lexika bereits festgestellt werden. Durch die Lemmatisierung entstehen gleiche Grundformen zu Wörtern die vormalig unterschiedliche Sentiment-Angaben hatten. Der Umstand zeigt jedoch, dass in der SA Probleme durch die Lemmatisierung entstehen können. Ferner können Lemmatisierer vor allem in flektionsreichen Sprachen wie dem Deutschen fehlerhafte Lemmas produzieren. Ein Hauptproblem der hier verwendeten Lemmatisierer ist, dass die Lemmatisierung häufig abhängig ist vom Satz in dem ein Wort steht und isolierte Wörter anders lemmatisiert werden als Wörter im Satzkontext. Dies kann zu fehlerhaften Wortformreduktionen und Problemen im Wort-Abgleich. Aus diesem Grund wird als mögliche Herangehensweise der SA auch eine ad-hoc-Verwendung von Lexika und Text ohne Lemmatisierung untersucht.

Als Standard-Methode wird der Text lemmatisiert, jedoch das Lexikon nicht. Dies ist das Verfahren, das man grundsätzlich als zielführend erachten würde, wenn in Lexika korrekte Grundformen enthalten sind und der Lemmatisierer die Grundformreduktion ebenso exakt durchführt. Beide Annahmen können jedoch nicht sicher belegt werden. Im Fall der Lemmatisierer konnte informell für einige Beispiele festgestellt werden, dass die Grundformreduktion fehlerhaft sein kann. Diese Standard-Methode wird dennoch in SA-Projekten für den Umgang mit Lemmatisierung umgesetzt (Kennedy & Inkpen, 2006; Taboada et al., 2011; Hogenboom et al., 2015).

Als letztes Verfahren wird die Lemmatisierung von sowohl Lexikon-Wörtern als auch Text mit anschließendem Wort-Abgleich für die SA. Dieses Verfahren kann sinnvoll sein um Eigenheiten in der Grundformreduktion zu standardisieren, so dass diese Eigenheiten sowohl im Text als auch im Lexikon implementiert sind und der Wort-Abgleich erfolgreich stattfindet.

In der vorliegenden Arbeit werden alle Verfahren systematisch evaluiert und auf Interaktionen mit den anderen Optionen der SA analysiert. Es stehen zwei verschieden Lemmatisierer zur Verfügung. Mehr zu diesen Lemmatisierern und zur Lemmatisierung auf Dramentextebene findet man in Kapitel 5.2.2.

5.3.4.2 Entwicklung

Der grundsätzliche Einsatz und die Entwicklung der Lemmatisierungskomponenten über die Programme lp_language_processor.py, lp_textblob_py und lp_treetagger wurde bereits in Kapitel 5.2.2 besonders in Bezug auf die Dramen-Text-Ebene erläutert.

Zentral für den Einsatz auf Lexikon-Ebene ist die Methode getLemma in der Textblob- und Treetagger-Klasse, über die das Lemma eines einzelnen Wortes ausgegeben wird. Die Probleme der Lemmatisierung eines einzelnen Wortes ohne korrekten Satzkontext wurden bereits angesprochen. Die lemmatisierten Versionen der einzelnen Lexika werden über Methoden in den jeweiligen Lexikon-Klassen und für Combined-Lexikon im Lexicon_Handler erstellt und ausgegeben (siehe Kapitel 5.3.1 und 5.3.2). Die Lemma-Versionen befinden sich als strukturierte txt-Dateien wie die herkömmlichen Versionen auch im Ordner SentimentAnalysis/TransformdedLexicons. Es wurden ebenso einsehbare lemmatisierte Versionen von DTA-erweiterten Lexika erstellt, die sich an gleicher Stelle befinden. Durch die Lemmatisierung von Lexikon-Einträgen wird das Lexikon im Normalfall verkürzt da unterschiedliche Flektionsformen auf gleiche Grundformen abgebildet werden. Als Beispiel sei das CD genannt, das in normaler Token-Form, nach Auflösen vorhandener Doppelungen und Entfernen neutraler Wörter, 7371 Einträge hat. Durch Lemmatisierung entstehen neue Doppelungen da manche Wörter auf die gleiche Grundform reduziert werden. Nach Auflösung dieser Doppelungen hat die Textblob-Version 7215 Einträge und die Treetagger Version 7093. Man erkennt, dass der Treetagger-Lemmatisierer mehr Wörter auf die gleiche Grundform reduziert. Dies gilt für alle Lexika.

5.4 Vorverarbeitung – Überblick

Über die Vorverarbeitung auf Text- und Lexikon-Ebene werden verschieden Optionen für Lexikon-basierte Verfahren in der vorliegenden Studie untersucht. Folgende Tabellen fassen alle in den Kapiteln 5.1 – 5.3 eingeführten Optionen zusammen mit den Benennungen, die in der weiteren Arbeit und im Code genutzt werden.

Tabelle 4: Verarbeitungs-Optionen – DTA-Erweiterung

Optionsname	DTAExtension	
Optionsausprägungen	noExtension	dtaExtended
Bedeutung	Keine Lexikonerweiterung über das DTA-Tool	Lexikonerweiterung über das DTA-Tool

Tabelle 5: Verarbeitungs-Optionen – Lemmatisierer

Optionsname	Lemmatizer		
Optionsausprägungen	tokens	textblob	treetagger
Bedeutung	Keine Lemmatisierung	Lemmatisierung über den pattern-Lemmatisierer der textblob-library	Lemmatisierung über den treetagger-Lemmatisierer

Tabelle 6: Verarbeitungs-Optionen – Lemmatisierungstyp

Optionsname	LemmatizationType		
Optionsausprägungen	noLemma	textLemma	bothLemma
Bedeutung	Keine Lemmatisierung	Lemmatisierung auf Text-Ebene	Lemmatisierung auf Text- und Lexikon-Ebene

Tabelle 7: Verarbeitungs-Optionen – Stoppwortlisten

Optionsname	Stopwords			
Optionsausprägungen	noStopwordList	standardList	enhancedList	enhancedFilteredList
Bedeutung	Keine Stoppwortliste	Eine herkömmliche Stoppwortliste	Die Standard-Liste erweitert mit den häufigsten Wörtern des Korpus	Die enhancedList jedoch ohne eindeutig Sentiment-tragende Wörter

Tabelle 8: Verarbeitungs-Optionen – Groß- und Kleinschreibung

Optionsname	CaseSensitivity	
Optionsausprägungen	caseInSensitive	caseSensitive
Bedeutung	Nicht-Beachtung der Groß- und Kleinschreibung im letzten Abgleichschritt	Beachtung der Groß- und Kleinschreibung im letzten Abgleichsschritt

5.5 Sentiment Analysis

In den folgenden Abschnitten wird die grundsätzliche Idee, Konzeption und tatsächliche programmiertechnische Umsetzung der Sentiment Analysis geschildert.

5.5.1 Konzeption

Es wird eine Pipeline aufgebaut bei der man aus den in den Kapiteln 5.1 – 5.3 formulierten Optionen zur Verarbeitung wählen kann: Lexikon, Lexikonerweiterung, Lemmatisierungsart, Lemmatisierer, Stoppwortliste und Beachtung von Groß- und Kleinschreibung im letzten Abgleichsschritt. Auf Basis dieser Optionen wird dann die SA durchgeführt.

Es werden SA-Metriken auf folgenden Ebenen berechnet:

- Drama
- Akt
- Szene
- Replik
- Sprecher (pro Drama, Akt, Szene, Replik)
- Sprecherbeziehung (pro Drama, Akt, Szene, Replik)

Zur Kalkulation wird für jede Ebene der dazugehörige Text gesammelt. Also für die Dramenebene der Gesamttext, für die Akt-Ebene nur der Text dieser Ebene usw. Die Drama-, Akt-, Szene- und Repliken-Ebene wird im Weiteren auch strukturelle Ebene genannt. Unter einem Sprecher wird eine Figur eines Dramas verstanden. Der dazugehörige Text sind alle Repliken die der Figur zugeordnet werden. Diese können wiederum pro Drama, Akt oder Szene betrachtet werden, sofern der Sprecher in der strukturellen Einheit auftritt. Zur Berechnung von Sprecherbeziehungen wird die Heuristik von Nalisnick und Baird (2013) herangezogen, die besagt, dass sich jeder Sprechakt

einer Figur auf den Sprecher der vorigen Replik bezieht. Die Heuristik geht also von einem ständigen Dialog zwischen Figuren aus, bei denen sich Sprechakte bezüglich ihres Sentiments auf die vorige Figur beziehen. Es ist naheliegend, dass diese Heuristik sehr fehlerbehaftet ist. Figuren stehen nicht immer im Dialog und beziehen sich in ihrer Replik nicht immer auf den Vorredner. Dennoch können Nalisnick und Baird (2013) mit der Heuristik erste Erfolge in der Literaturanalyse erzielen, weswegen die Idee, auch aufgrund seiner Einfachheit in der Umsetzung, im vorliegenden Projekt aufgegriffen wird um die Tauglichkeit und den Nutzen dieser zu untersuchen. Für die Sprecherbeziehungen werden also für alle Ausgangssprecher jeweils alle Repliken gesammelt, die sich auf einen jeweiligen Zielsprecher als Vorredner beziehen. Auch dieses Konzept kann für das Gesamtdrama, den Akt oder die Szene betrachtet, insofern eine Sprecherbeziehung nach dieser Heuristik für eine strukturelle Einheit zu Stande kommt.

Für die SA wurden Programme zur Modellierung, Kalkulation, Durchführung und Ausgabe entwickelt, die noch in Kapitel 5.5.3 genauer erläutert werden. Die Grund-Konzeption beruht darauf, nach Vorverarbeitung des Textes und Einstellung des Lexikons über die SA-Optionen, im jeweiligen Dramentext SBWs durch Abgleich der Wörter des Textes mit den Lexika zu identifizieren und diese mit den Sentiment-Ausprägungen zu sichern. Dann werden durch simple Kalkulationsverfahren die SA-Metriken berechnet. Für verschieden Lexika können dabei mehrere Metriken berechnet werden, wie im nächsten Kapitel noch erläutert wird.

5.5.2 SA-Metriken

Für die oben definierten Ebenen werden verschieden Metriken berechnet, für manche Lexika auch mehrere. Es wird grundsätzlich zwischen zwei Konzepten unterschieden. Zum einen die Berechnung mittels Term-Zähl-Verfahren (Turney, 2002; Kennedy & Inkpen, 2006), bei der die Zahl an Wörtern bestimmter Sentiment-Klassen den Endwert bestimmt und die Kalkulation über Polaritäts-Stärken, die ähnlich verläuft, wobei jedoch nicht die Anzahl der Terme entscheidend ist, sondern die Sentiment-Ausprägung in Form von Polaritätsstärken. Ferner wird für den Bereich der Polaritäts-Sentiments zwischen der Positivität, der Negativität und der Polarität unterschieden. Die Positivität und Negativität ist das Ergebnis durch Kalkulation bezüglich dieser Sentiment-Klassen, die Polarität wird berechnet durch Abzug der Negativität von der Positivität

und gibt die Gesamtausrichtung eines Textes im Verhältnis dieser beiden Klassen wider. Die Polarität wird auch genutzt, um für eine Untersuchungseinheit insgesamt anzugeben, ob diese positiv oder negativ ist. Ist die Polarität kleiner 0 ist die Einheit negativ, ist sie größer 0 positiv.

Zur Illustration wird ein fiktives Beispiel einer Replik angenommen, bezüglich der für ein Lexikon festgestellt wurde, dass es drei negative und zwei positive Wörter enthält. Über Term-Zähl-Methodik erhält die Positivität den Wert 2, die Negativität 3 und die Polarität -1. Die Replik wird über diese Methodik als negativ angenommen. Geht man nun davon aus, dass die drei negativen Wörter Polaritätsstärken von 0,1, 0,3 und 0,5 während die positiven Wörter 0,5 und 0,9 haben erhält man über die Polaritätsstärken einen Wert von 0,9 für die Negativität und von 1,4 für die Positivität, was wiederum eine Polarität von +0,5 ergibt und somit auf eine positiv konnotierte Replike hindeutet. Metriken die über Term-Zähl-Methodik berechnet werden, werden im Folgenden auch dichotome Metriken genannt, da sie wie dichotome Gewichte aufgefasst werden können, die nur die Werte 0 oder 1 annehmen können, um zum gleichen Kalkulationsergebnis zu kommen.

Für jedes Lexikon existiert zumindest eine Metrik gemäß Term-Zähl-Verfahren. Bei Lexika mit Polaritätsgewichten wird eine Metrik über die Polaritätsgewichte berechnet und eine über Term-Zähl-Verfahren. Die Polaritätsgewichte werden hierzu dichotomisiert auf eine 1 für die Positivität bei einem positiven Gewicht oder einem Gewicht, das der positiven Klasse zugewiesen ist und eine 1 für die Negativität bei eine negativen Gewicht. Dies ist dann äquivalent zum Term-Zähl-Verfahren ohne Beachtung der Gewichte. Folgende Tabellen fassen alle Metriken pro Lexikon zusammen. Es wird der Name angegeben, der im Folgenden und im Code genutzt wird sowie das Berechnungsverfahren. Liegen für ein Lexikon sowohl Metriken für Polaritätsangaben als auch dichotome vor, werden die dichotome durch ein „Dichotom“ im Namen verdeutlicht. Gibt es ohnehin nur eine Methodik wird meist der herkömmliche Lexikonname und das Sentiment verwendet.

Zunächst die Metriken, die sich auf das Lexikon SentiWS beziehen.

Tabelle 9: Sentiment-Metriken SentiWS

Name	Methodik
positiveSentiWS	Addition von Polaritätsstärken
negativeSentiWS	Addition von Polaritätsstärken
polaritySentiWS	positiveSentiWS - negativeSentiWS
positiveSentiWSDichotom	Term-Zähl-Verfahren
negativeSentiWSDichotom	Term-Zähl-Verfahren
polaritySentiWSDichotom	positiveSentiWSDichotom - negativeSentiWSDichotom

Auch die Metriken zum BAWL-R enthalten dichotome Versionen und solche mit Polaritätsstärken. Aufgrund der Benennung im Original-BAWL-R wird die Polarität über Polaritätsstärken emotion genannt.

Tabelle 10: Sentiment-Metriken BAWL-R

Name	Methodik
positiveBawl	Addition von Polaritätsstärken
negativeBawl	Addition von Polaritätsstärken
emotion (polarityBawl)	positiveBawl - negativeBawl
positiveBawlDichotom	Term-Zähl-Verfahren
negativeBawlDichotom	Term-Zähl-Verfahren
polarityBawlDichotom	positiveBawlDichotom - negativeBawlDichotom

Für das NRC liegen nun neben den dichotomen Polaritätsvariablen auch noch Emotionskategorien vor, die über ein herkömmliches Term-Zähl-Verfahren kalkuliert werden. Ferner wird noch eine Metrik emotionPresent definiert, die alle vorhandenen Emotionswörter addiert. Diese Metrik ist nicht unbedingt gleich der Summe der einzelnen Emotionswerte, da im NRC ein Wort durchaus mehrere Emotionsannotationen haben kann (siehe auch Kapitel 2.3.3).

Tabelle 11: Sentiment-Metriken NRC

Name	Methodik
positiveNrc	Term-Zähl-Verfahren
negativeNrc	Term-Zähl-Verfahren
polarityNrc	positiveNrc – negativeNrc
anger	Term-Zähl-Verfahren
anticipation	Term-Zähl-Verfahren
disgust	Term-Zähl-Verfahren
fear	Term-Zähl-Verfahren
joy	Term-Zähl-Verfahren
sadness	Term-Zähl-Verfahren
surprise	Term-Zähl-Verfahren
trust	Term-Zähl-Verfahren
emotionPresent	Term-Zähl-Verfahren

Die Metriken für das CD sind wieder wie gewöhnlich. Es sei hierbei jedoch zu beachten, dass der Unterschied zwischen der dichotomen Metrik und derjenigen mit Polaritätsstärken sehr gering ist, da die Polaritätsstärken hier nur Ausprägungen mit den Werten 0,7 und 1 haben können.

Tabelle 12: Sentiment-Metriken CD

Name	Methodik
positiveCD	Addition von Polaritätsstärken
negativeCD	Addition von Polaritätsstärken
polarityCD	positiveCD – negativeCD
positiveCDDichotom	Term-Zähl-Verfahren
negativeCDDichotom	Term-Zähl-Verfahren
polarityCDDichotom	positiveCDDichotom - negativeCDDichotom

Für das GPC gibt es nur eine dichotome Ausprägung mit Berechnung über Term-Zähl-Verfahren:

Tabelle 13: Sentiment-Metriken GPC

Name	Methodik
positiveGpc	Term-Zähl-Verfahren
negativeGpc	Term-Zähl-Verfahren
polarityGpc	positiveGpc - negativeGpc

Für das kombinierte Lexikon wurden spezielle Metriken berechnet. Die Art der Lexikon-Kombination und die Kalkulation wurden bereits in Kapitel 5.3.2 angesprochen. Für positiveCombined und negativeCombined wird vereinfach erklärt pro Wort betrachtet, ob für irgendein Wort aller Lexika eine Polarität vorliegt und die Gesamtwerte gemäß Term-Zähl-Methodik berechnet. Bei Ambiguitäten wird eine Mehrheitsentscheidung durchgeführt, das heißt die Polarität gewählt, die in den meisten Lexika vorkommt. Bei Unentschieden entscheidet dasjenige Lexikon, dass gemäß der in Kapitel 8 durchgeführter Evaluation die beste Erkennungsleistung aufbringt. Für clearlyPositiveCombined und clearlyNegativeCombined wird ein Wort nur in das Term-Zähl-Verfahren aufgenommen, wenn es in mindestens 3 Lexika (also der Mehrheit) mit einer konsistenten Annotation vorkommt. Die Polaritätsangabe wird dann als eindeutig belegt betrachtet.

Tabelle 14: Sentiment-Metriken Kombiniertes Lexikon

Name	Methodik
positiveCombined	Term-Zähl-Verfahren
negativeCombined	Term-Zähl-Verfahren
polarityCombined	positiveCombined - negativeCombined
clearlyPositiveCombined	Term-Zähl-Verfahren
clearlyNegativeCombined	Term-Zähl-Verfahren
clearlyPolarityCombined	clearlyPositiveCombined - clearlyNegativeCombined

Alle genannten Metriken können absolut und normalisiert betrachtet werden. Mohammad (2011) exploriert den Nutzen von Normalisierungen. Absolute Werte geben die Sentiment-Ausprägung unabhängig von der Länge des Textes an. Es ist naheliegend, dass in vielen Fällen die Länge eines Textes Einfluss auf die Werte hat, vor allem

bei Term-Zähl-Verfahren. Längere Texte können eher starke Ausprägungen haben, als weniger lange Texte. Für Polaritätswerte ist dieser Einfluss nicht so stark, da längere Texte größere Chancen haben sowohl positive als auch negative Sentiments zu enthalten, weswegen in vielen Fällen die Polaritätswerte von langen Texten auch mit kurzen vergleichbar sind. Für die anderen Werte, z.B. die Emotions-Metriken ist jedoch der Einfluss längerer Texte auf die Ausprägung klar, da längere Texte mehr Emotionswörter enthalten können. Die absoluten Werte können durchaus sinnvoll für die Analyse sein; man kann argumentieren, dass eine Replik, die länger ist und dadurch mehr Zorn-Wörter enthält insgesamt tatsächlich eine zornigere Replik ist, als eine kürzere Replik mit einem geringeren Zorn-Wert. Es sind jedoch Anwendungsfälle denkbar bei denen diese Werte vergleichbarer betrachtet werden wollen, unabhängig von der Länge. Für diesen Fall wird im vorliegenden Projekt Normalisierung eingesetzt. Mohammad (2011) spricht in seinem Projekt von emotionaler Dichte.

In der vorliegenden werden zwei Normalisierungsarten verwendet. In beiden Fällen werden die absoluten Werte durch eine bestimmte Zahl geteilt, um die Werte unabhängig von der Länge vergleichbar zu machen. Im ersten Fall findet eine Normalisierung an der Zahl aller dazugehörigen SBWs des Textes der zu analysierenden Einheit statt. Unter den SBWs für diese Form der Normalisierung werden alle SBWs verstanden, die zu der betrachteten Metrik und dem Sentiment-Lexikon gehören, also im Fall von polaritySentiWSDichotom alle Wörter, die entweder als positiv oder negativ bezüglich SentiWS ausgezeichnet sind. Beim NRC werden für diese Normalisierung bei der Polarität auch nur diejenigen Wörter verwendet, die entweder negativ oder positiv annotiert sind. Für die Emotionskategorien nur Wörter die emotionsbeladen gemäß NRC sind. Für alle dichotome Metriken außer die Polarität gibt die Normalisierung an der Zahl an SBWs den Anteil der einzelnen Klasse an, also im Fall von 10 Polaritätswörtern und 3 davon sind positiv ergibt die Normalisierung der Positivitätklasse 0,3 also einen Anteil von 30%. Für die Polarität und Maße mit Sentiment-Stärken ist diese Aussage mathematisch nicht gültig. Nichtsdestotrotz ermöglicht die Normalisierung in diesem Fall die bessere Vergleichbarkeit unterschiedlich langer Texte.

Als zweite Normalisierungsform wird noch die Normalisierung an der Gesamtanzahl der Worte implementiert. Dabei wird der absolute Wert durch die Zahl aller Wörter geteilt. Im Fall von dichotomen Metriken außer der Polarität erhält man dadurch

den Anteil der jeweiligen Sentiment-Klasse am Gesamttext. Beispielsweise im obigen Fall, bei 3 positiven Wörtern in einem Text von 20 Wörtern ergibt das einen normalisierten Wert von 0,15 also 15% positive Wörter im Text. Auch hier gilt die gleiche Annahme aus mathematischen Gründen nicht für Polaritäten und generell nicht für alle Maße basierend auf Gewichten. Es kann aber auch in dem Fall die Vergleichbarkeit von Werten unabhängig von der Länge besser ermöglicht werden.

Folgende Tabelle fasst alle Normalisierungsmöglichkeiten zusammen. Es werden ferner die Listennamen angegeben, die auch tatsächlich im Code zur Speicherung der Metriken verwendet werden.

Tabelle 15: Sentiment-Metriken Normalisierungen

Normalisierungsliste	Normalisierungsart
metricsTotal	Der absolute Werte ohne Normalisierung
metricsNormalisedSBWs	Normalisiert an der Zahl der SBWs
metricsNormalisedLengthInWords	Normalisiert an der Zahl aller Wörter

Obschon die Normalisierung von absoluten Werten in den meisten Fällen die Vergleichbarkeit steigert, können verschiedene mathematische Effekte auftreten die stets differenziert analysiert werden müssen. Häufig können sehr kurze Texte sehr hohe normalisierte Werte erreichen, weil sie beispielsweise nur aus einen Sentiment-tragenden Wort bestehen, was zu eine maximal hohen normalisierten Wert führt. Die Interpretation von Werten absolut oder normalisiert kann für verschiedene Anwendungsfälle also sehr komplex sein und unterschiedlich ausfallen, die Art der Metrik und der Normalisierung sowie die untersuchte Einheit und der potentielle Vergleich, den man unternimmt, müssen stets beachtet werden. Über die beschriebene differenzierte Betrachtung verschiedener Lexika und verschiedener Berechnungsverfahren können jedoch insgesamt gewinnbringende Erkenntnisse über den Einsatz von Lexikon-basierten Verfahren gewonnen werden.

5.5.3 Entwicklung

Bei der Entwicklung der Sentiment Analysis wird zwischen vier größeren Einheiten unterschieden: Verschiedene SA-Klassen zur Modellierung und strukturierten Speicherung von SA-Konzeption in der Datei sa_models, die Kalkulationseinheit im Pro-

gramm sa_calculator, die Verarbeitungs- und Durchführungsklasse in der Datei sa_sentiment_analysis die alle Bestanteile zusammenführt und die SA durchführt sowie die Ausgabe der berechneten SA-Daten im Programm sa_output.py. Die komplette Zusammensetzung und Funktionalität wird im Folgenden vereinfacht dargestellt. Zur detaillierteren Einsicht wird auf den Programm-Code verwiesen.

Es ist an dieser Stelle wichtig mitzuteilen, dass die gesamte SA immer für alle Metriken gleichzeitig durchgeführt wird, d.h. eine konkrete Auswahl eines Lexikons findet nicht statt, es werden immer alle Metriken berechnet. Erst in der Ausgabe kann bestimmt werden welche Daten genau eingesehen werden wollen. Es hat sich herausgestellt, dass die Kalkulation keinen Einfluss auf die Performanz hat und es einfacher für die spätere Gestaltung der Ausgabe und für die Analyse ist, gleich und immer alle Metriken zu berechnen. Dazu wird das Lexikon CombinedLexicon initialisiert und genutzt. Es handelt sich um eine Datenstruktur, die die Wörter aller Lexika mit allen dazugehörigen Lexikon-spezifischen Sentiment-Informationen speichert (siehe Kapitel 5.3.2.2).

Die zentrale Methode zur Durchführung der SA befindet sich in der Klasse Sentiment_Analyzer in der Datei sa_sentiment_analysis und lautet attachAllSentimentInfoToDrama. Der Klasse werden alle Optionen für die SA übergeben, also ob man ein erweitertes Lexikon verwenden soll, der Lemmatisierer, der Lemmatisierungstyp, die Stoppwortliste und ob Groß- und Kleinschreibung im letzten Abgleichschritt beachtet werden sollen. Die einzelnen Optionen werden dann bei allen Kalkulations- und Verarbeitungsschritten beachtet. Wie erwähnt wird nicht angegeben, welches Lexikon man genau betrachtet, es werden immer alle Metriken gleichzeitig berechnet. Der Methode wird ein über den Pre_Processor vorverarbeitetes Dramen-Modell übergeben (siehe Kapitel 5.2.1).

Der grundlegende erste Schritt für die SA ist das Anfügen von SBWs an die einzelnen Drameneinheiten wie sie im vorverarbeiteten Dramenmodell bereits vorliegen. Ein SBW wird in sa_models als Klasse Sentiment_Bearing_Word modelliert. Ein SBW besteht aus Lemma, Token, POS und den Sentiment-Werten für alle Metriken. Im Sentiment_Analyzer findet gemäß gesetzter Optionen ein Abgleich des Dramentextes mit den Lexikon-Wörtern statt. Ist ein Wort bzw. Lemma (je nach ausgewählter Option) im Lexikon CombinedLexicon enthalten, so werden die Attribute des SBWs gesetzt. Sen-

timent-Werte von Metriken für Lexika in denen das Wort/Lemma nicht enthalten ist, werden mit 0 gesetzt. Diese SBWs werden den einzelnen Dramenmodell-Einheiten (Drama, Akt, Szene usw.) für die spätere Kalkulation hinzugefügt. Selbiges Verfahren wird für die Sprecher und die Sprecherbeziehungen durchgeführt.

Für die Sprecherbeziehungen werden diese zuvor erst noch modelliert und die dazugehörigen Repliken hinzugefügt. Dafür wird in den Modellen die Klasse Sentiment_Relation definiert, die eine solche Charakter-zu-Charakter-Beziehung modelliert. Der Klasse wird bei der Initialisierung der Ausgangs- und Zielsprecher sowie die dazugehörigen Repliken gemäß der gewählten Heuristik übergeben (siehe auch Kapitel 5.5.1). Derartige Modelle werden in der Untermethode attachSentimentRelationsToSpeaker der Methode attachAllSentimentInfoToDrama konstruiert und den Sprecher-Modellen des Dramenmodells hinzugefügt.

Die finale Kalkulation aller absoluter und normalisierter Metriken wird über die Klasse Sentiment_Calculator durchgeführt, die auf Basis der SBWs und weiterer Längenangaben der zu betrachtenden Einheit alle Metriken kalkuliert und als Objekt der Klasse Sentiment_Metrics zurückgibt. Hierbei handelt es sich um eine Speicher- und Ausgabestruktur für alle finalen Sentiment-Einheiten. Diese Struktur wird den herkömmlichen Dramen-Einheiten im Attribut _sentimentMetrics von außen zugreifbar angefügt. Auf die beschriebene Art werden alle Sentiment-Metriken für die angegebenen Optionen für alle definierten Ebenen kalkuliert.

Über ein Ausgabe-Programm kann man sich die Ergebnisse aufbereitet in verschiedenen Formaten ausgeben lassen. Es handelt sich um das Programm sa_output, in der die Klasse Sentiment_Output_Generator definiert wird, über welche verschiedenen Methoden zur Kalkulation und Ausgabe bereit gestellt sind. Es gibt Methoden, um den Output für ein einzelnes Drama oder mehrere zu generieren. Je nachdem muss der Pfad zur Pickle-Datei des erwünschten Dramas oder zum Ordnerpfad der Dramensammlung übergeben werden. Die vorverarbeiteten Dramen werden durch die Klasse Pre_Processing erstellt und liegen im Ordner Python/Dumps/ProcessedDramas (siehe Kapitel 5.2.1). Den Ausgabemethoden werden ferner die SA-Optionen für die Durchführung der SA und die Initialisierung des Sentiment_Analyzer-Objektes übergeben. In den jeweiligen Ausgabe-Methoden werden die Dramen eingelesen und die SA über ein Objekt der Klasse Sentiment_Analyzer durchgeführt.

Die Methoden generateJSONFileForAllDramas und generateJSONFileForSingleDrama generieren eine strukturierte JSON-Datei in der alle Sentiment-Daten sowie einige Meta-Daten gemäß Ebene gesichert sind. Diese wird auch für das Front-End genutzt. Die Methoden processAndCreateTxtOutputMutlipleDramas und createTxtOutputSingleDrama kreieren einen strukturierten Txt-Output für die Metriken. Den einzelnen Methoden wird auch ein Ausgabepfad mit Dateinanme übergeben. Einige Beispielausgaben befinden sich im Ordner SentimentAnalysis/SA-Output.

6 Vokabular-basierte Evaluation

Als erster Zwischenschritt zur Evaluation der Lexika und einiger Verfahren aus Kapitel 5 wurde eine Evaluation durchgeführt, die darauf basiert, abzugleichen, wie hoch der Anteil erkannter Wörter eines Lexikons am Gesamtkorpus ist. Dieses Verfahren wird im Folgenden Vokabular-basierte Evaluation genannt. Die genaue Idee und Entwicklung wird in diesem Kapitel beschrieben. Die Ergebnisse werden lediglich knapp zusammengefasst und auf den Anhang verwiesen, da die Gold-Standard-Evaluation (siehe Kapitel 8) als etablierte Methodik in der Forschung, die vorrangige Evaluationsmethodik in dieser Arbeit ist.

6.1 Idee und Vorgehen

Die grundsätzliche Motivation für die Vokabular-basierte Evaluation ist eine erste informelle Analyse des Korpus-Wortschatz und ein Vergleich dieses Wortschatzes mit dem Wortschatz der normalen Lexika und der verarbeiteten (erweiterten, lemmatisierten) Lexika. Es wird auf Basis des veralteten und poetischen Wortschatzes die Annahme aufgestellt, dass die Wort-Einträge der Lexika stärker vom Wortschatz des Korpus abweichen als dies bei sonstigen Untersuchungsgegenständen in der SA, wie z.B. Produkt- und Film-Reviews, der Fall ist. Dies betrifft den grundsätzlichen Wortschatz aber auch die Orthographie wie z.B. das häufig im Korpus verwendete Wort „betriegen“ statt betrügen zeigt. Diese Annahme ist naheliegend da die verwendeten SA-Lexika auf modernen Online-Lexika (Vo et al., 2009; Clematide und Klenner, 2010), der Analyse von Produkt-Review-Korpora (Remus et al., 2010) und News-Korpora (Clematide und Klenner, 2010) oder Crowdsourcing (Mohammad & Turney, 2010), basieren. Über die vokabular-basierte Evaluation soll durch einen prozentualen Wortabgleich der

Wörter eines Lexikons mit dem Wortschatz des Korpus in Erfahrung gebracht werden, welches Lexikon und welche Methode die meisten Wörter als SBWs im Test-Korpus erkennt. Es handelt sich also um eine Evaluation auf der reinen Wort-Ebene ohne tatsächliches Sentiments zu betrachten. Es kann auf Basis dieser Evaluation keine Aussage darüber getroffen werden, ob die SA der einzelnen Lexika und Methode tatsächlich im Vergleich korrekter funktioniert, da eine erhöhte Zahl erkannter Wörter nicht bedeutet, dass diese bezüglich ihres Sentiments in den Texten korrekt erkannt wurden. Waltinger (2010) beispielsweise stellt fest, dass größere Lexika durchaus eine schlechtere Erkennungsleistung als kleinere haben können, da die Genauigkeit in der Sentiment-Erkennung leiden kann, wenn zu übermäßig viele Wörter ungenau als Sentiment-tragend ausgezeichnet werden. Die Vokabular-Evaluation wird aufgrund dessen auch lediglich als erste informelle Orientierung betrachtet. Ähnliche Verfahren in der Forschung sind nicht bekannt. Es können deswegen auch keine Vergleichswerte herangezogen werden.

Zur Vereinfachung werden als Text- und Lexikonverarbeitungsschritte nur die DTA-Erweiterung, der Lemmatisierungstyp und der Lemmatisierer analysiert. Bezüglich der Stoppwörter wird nur zwischen keinen und dem Einsatz der Standardliste unterschieden (noStopwords, standardList). Als letzter Abgleichsschritt wird die Groß- und Kleinschreibung nicht beachtet (caseInSensitive). Diese Vereinfachung wurde durchgeführt da weniger die einzelnen SA-Optionen analysiert und evaluiert werden sollen, sondern die Lexika mit ihrem Wortschatz.

Über verschiedene Programme wird zunächst der Wortschatz des ganzen Korpus und der einzelnen Dramen akquiriert und aufbereitet. Dieser Wortschatz wird am weiteren auch als das Vokabular bezeichnet. Es handelt sich um Wortlisten mit Angaben darüber, wie häufig einzelne Wörter vorkommen. Diese Wortlisten wurden für alle in der Evaluation implementierten SA-Optionen vor, also mit und ohne Stoppwörter, sowie lemmatisiert und nicht lemmatisiert. Die Wortlisten sind nach Häufigkeit der Wörter geordnet. Sie können in zukünftigen Projekten auch verwendet werden, um den Wortschatz Lessings im Allgemeinen zu analysieren. Das entwickelte Framework zur Erstellung dieser Dateien kann auch für andere Dramen zur Wortschatzanalyse genutzt werden. Die Entwicklung wird noch im nächsten Kapitel genauer beschrieben. Sie befinden sich im Ordner Word-Frequencies im Anhang. Es wird dann simpel abge-

glichen zu welchem Anteil Wörter der verarbeiteten Lexika im Vokabular erkannt werden. Auf Basis der verschiedenen Verarbeitung auf Lexikon- und Vokabular-Ebene wird nicht nur untersucht welche Lexika mehr Vokabular-Wörter enthalten sondern auch welche der zentralen Methode zu einem erhöhten Wortschatz-Anteil führt, also DTA-Erweiterung (noExtension, DTAExtension), Lemmatisierungstyp (noLemma, textLemma, bothLemma) und Lemmatisierer (textblob/pattern, treetagger). Es werden Ausgabedateien produziert mit verschiedenen Vokabular-, Lexikon- und Wortanteil-Informationen. Zentral ist dabei der Prozentsatz der verschiedenen erkannten Wörter und der erkannten Wörter insgesamt. Ferner werden auch die erkannten Wörter ausgegeben. Alle Analysen wurden auf dem Gesamtkorpus und pro Drama durchgeführt, um auch dramenspezifische Besonderheiten zu identifizieren.

Über die beschriebene Vokabular-basierte Evaluation soll ein erster Einblick auf mögliche Leistungsunterschiede von Methoden und Lexika gewonnen werden und die detaillierte Analyse des Wortschatzes ermöglicht werden.

6.2 Entwicklung

Im folgenden Abschnitt werden die Programm-Dateien für die Vokabular-basierte Evaluation grob beschrieben. Für genauere Einsicht wird auf den Code im Anhang verwiesen.

Über die lp_dramaLanguage_output.py wird die Klasse DramaLanguage_Output implementiert, die Methoden zur Verarbeitung und Produktion von Vokabular-Listen der Dramentexte zur Verfügung stellt. Je nachdem welche Verarbeitungsform man anstrebt werden Namen von Lemmatisierern und Stoppwortlisten übergeben. In der vorliegenden Evaluation wurden aber lediglich Vokabular-Listen mit und ohne Stoppwörter über die Stoppwortliste standardList betrachtet. Mittels verschiedener Methoden lässt sich eine Vokabular-Liste für den ganzen Korpus oder einzelne Dramen produzieren. Die Methoden lesen die Dramen aus dem Ordner Lessing-Dramen/ ein. Zur Sprachverarbeitung werden Objekte der Klasse Language_Processor verwendet.

Bei den Ausgabedateien handelt es sich um strukturierte txt-Dateien, die neben Metadaten wie die Anzahl aller Wörter insgesamt und die Anzahl aller unterscheidbarer Wörter alle Wörter des Dramas oder Gesamtkorpus nach Häufigkeit geordnet auf-

listen. Die Häufigkeit wird mitangegeben. Bei lemmatisierten Versionen wird das Lemma, die POS sowie alle dazugehörigen Tokens pro Zeile angegeben. Anbei ein Ausschnitt der Stoppwort-befreiten (standardList) Liste für das Gesamtkorpus lemmatisiert mit treetagger:

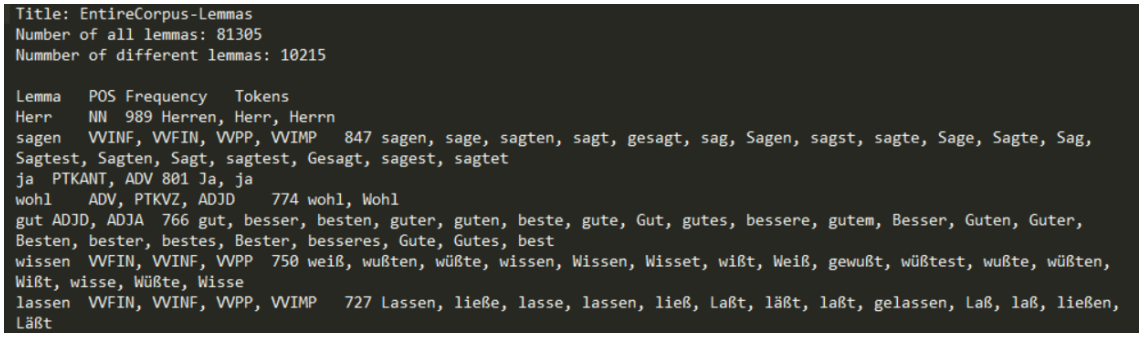


Abbildung 7: Ausschnitt Gesamt-Korpus Vokabular lemmatisiert

Die Dateien befinden sich im Ordner Word-Frequencies im Anhang.

In der Datei evaluation_vocabulary.py nun wird die tatsächliche Vokabular-basierte Evaluation umgesetzt. Über die Klassen Vocabulary und Evaluation_Result_Vocabulary werden Methoden und Speicherstrukturen zur Modellierung eines Text-Vokabulars und eines Evaluationsergebnisses implementiert. Über die Klasse Evaluation_LexiconVsVocabulary wird die tatsächliche Evaluation umgesetzt. Es wurden Methoden zur Produktion für Evaluationsdateien für einzelne oder mehrere Lexika und einzelne oder mehrere Vokabular-Dateien entwickelt. Je nach Methodik müssen Pfade für die Vokabular-Dateien und Namen der Lexika angegeben werden. Methoden, die die automatische Generierung von allen möglichen Kombinationen implementieren, gehen für das Einlesen und Ausgeben vom Standardpfaden aus, also für Vokabular-Dateien der Ordner Word-Frequencies/. Die Lexikon-Namen können wie in der Datei lexicon_handler.py definiert gewählt werden.

Die erzeugten Evaluationsdateien geben die Länge des Lexikons, die Länge des gewählten Vokabulars (verschiedene Wörter), die erkannten unterscheidbaren Wörter, den Anteil der erkannten unterscheidbaren Wörter, die Länge des Vokabulars insgesamt, die erkannten Wörter insgesamt und den Anteil der erkannten Wörter insgesamt an. Ferner werden alle erkannten Wörter aufgelistet. Folgender Screenshot zeigt als Beispiel einen Abschnitt der Ausgabe für das BAWL-R über die Methodik (noExtension, tokens, noLemmas):

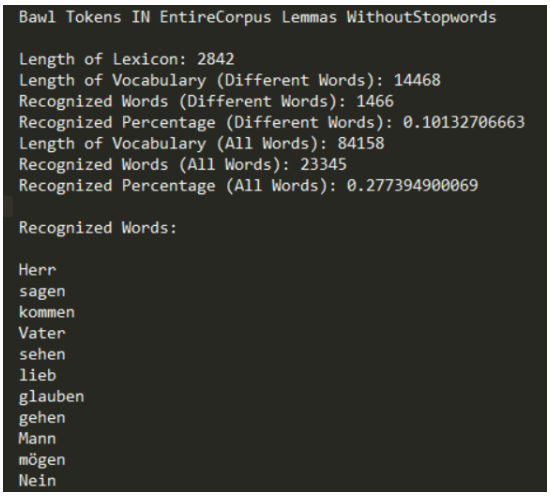


Abbildung 8: Ausschnitt Vokabular-basierte Evaluationsdatei

Alle Evaluationsdateien für alle Verfahren und Lexika befinden sich zur detaillierten Einsicht im Ordner Evaluation/Vocabulary-Evaluation. Die Ordnerstruktur orientiert sich nach Lexika und gewählter Methodik.

6.3 Ergebnisse

Es wurden für eine differenzierte Analyse oben beschriebene Evaluationsdateien für alle kombinatorischen Möglichkeiten an Verfahren und für jedes Einzeldrama und den Gesamtkorpus erstellt. Die erhobenen Ergebnisse der Vokabular-basierten Evaluation wurden analysiert und aufbereitet. Die wichtigsten Ergebnisse wurden in Tabellen (Google Tables)[4] zusammengefasst, die im Anhang einsehbar sind.

Im Folgenden werden die zentralen Ergebnisse und Erkenntnisse knapp und verkürzt beschrieben, da die Vokabular-basierte Evaluation ein sekundäres Evaluationsverfahren zur ersten informellen Analyse ist. Die Grafiken wurden mit Hilfe von Google Tables erstellt. Die nachfolgenden Beschreibungen beschränken sich auf den Gesamtkorpus als Wort-Vokabular für den Abgleich. Ferner beschränkt man sich auch noch auf das Stoppwort-bereinigte Vokabular. Als zentrale betrachtete Werte werden im Folgenden der Anteil erkannter unterscheidbarer Worte an allen unterscheidbaren Worten sowie der Anteil aller erkannter Wörter insgesamt an allen Wörtern genutzt. Dramenspezifische Ergebnisse sowie Resultate bezüglich Stoppwörtern werden weiter unten kurz zusammengefasst.

Folgendes erste Diagramm zeigt die Wort-basierte Erkennungsrate der Lexika ohne größere Bearbeitung, also ohne Lexikonerweiterung und Lemmatisierung:

[4] https://www.google.com/sheets/about/

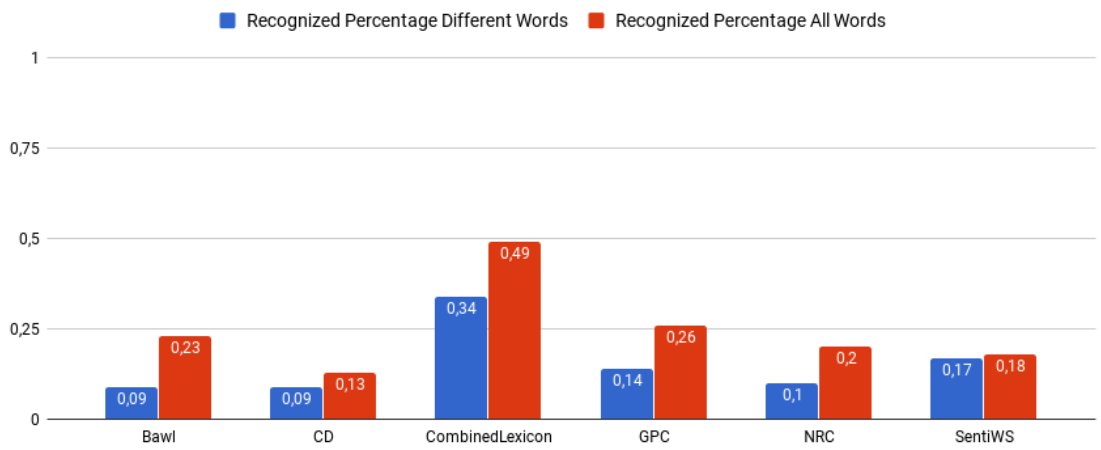


Abbildung 9: Vokabular-basierte Evaluation – Balkendiagramm – Lexikonvergleich

Man kann trivialerweise feststellen, dass das kombinierte Lexikon, die meisten Token des Korpus erkennt, da es alle Lexika zusammenfasst. Insgesamt erkennen also alle Lexika zusammen etwa 34% verschiedene Wörter im Korpus als Sentiment-Tragend. Alle Wörter betrachtend wird somit sogar fast die Hälfte des Gesamtkorpus als SBW identifiziert, d.h. viele der unterscheidbaren Wörter kommen übermäßig häufig vor. Bezüglich der Einzellexika fällt auf, dass SentiWS die beste Erkennungsrate unterscheidbarer Wörter aufweist und diese sich kaum zur Erkennungsrate aller Wörter unterscheidet, d.h. SentiWS enthält viele eher seltene Wörter. Die Lexika NRC, Bawl und CD erkennen lediglich 10% unterscheidbare Wörter, haben aber eine deutlich höhere Erkennungsrate bei der Betrachtung aller Wörter. Sie enthalten also viele häufige Wörter wie möglicherweise Stoppwörter, die die Stoppwortliste nicht erfasst.

Bezüglich der Sprachverarbeitung wird lediglich der Anwendungsfall treetagger und bothLemma betrachtet. Also die Nutzung des treetagger-Lemmatisierers und Lemmatisierung auf Lexikon- und Textebene. Die Ergebnisse waren für diesen Fall für die meisten Lexika am besten. Sonstige Besonderheiten bezüglich Lemmatisierung werden noch weiter unten kurz angesprochen. Folgende Ergebnisse beziehen sich auf die nicht DTA-erweiterten Lexikon-Versionen.

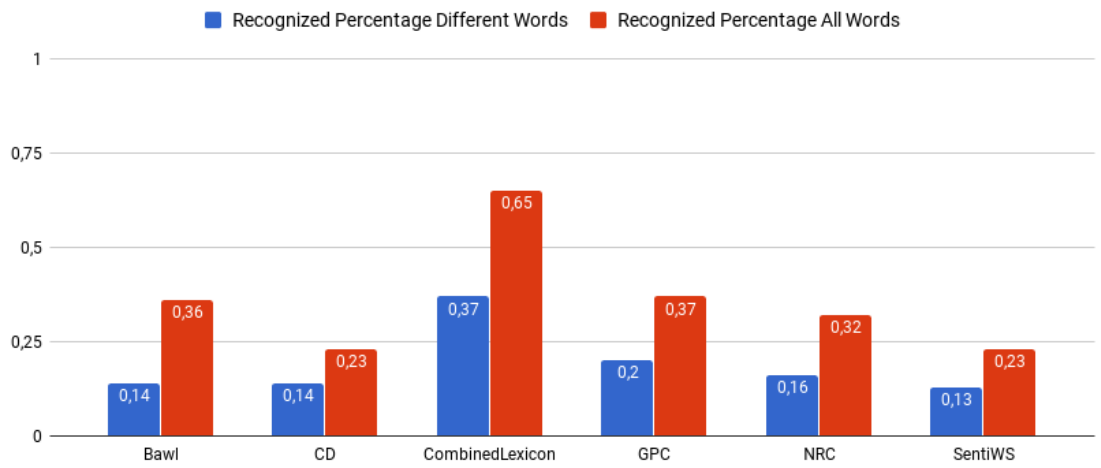


Abbildung 10: Vokabular-basierte Evaluation – Balkendiagramm – Lexikonvergleich mit DTA-Erweiterung

Man kann für alle Lexika eine grundsätzliche Steigerung bezüglich der Erkennungsrate von Wörtern durch Lemmatisierung feststellen. Ausnahme dafür ist dabei das Lexikon SentiWS, das zahlreiche flektierte Formen von Wörtern enthält. Für SentiWS wird die Erkennungsrate für unterscheidbare Wörter schlechter. Dies kann an Problemen in der Grundformreduktion des Lemmatisierers liegen insofern, dass die flektierten Formen in Lexikon und Text nicht korrekt lemmatisiert werden und die zuvor auffindbaren Flektionsformen aufgrund dieser Probleme nicht mehr identifiziert werden. Im Vergleich kann man festhalten, dass von allen Einzellexika das GPC nun die beste Erkennungsrate bezüglich unterscheidbarer aber auch aller Wörter insgesamt hat.

Eine deutliche Leistungssteigerung bezüglich der Erkennungsrate wird durch die DTA-Erweiterung mit historischen linguistischen Varianten konstatiert. Als Beispiel wird hierzu lediglich die Roh-Form der Lexika und des Textes analysiert also die DTA-erweiterten Lexika ohne Einsatz von Lemmatisierung.

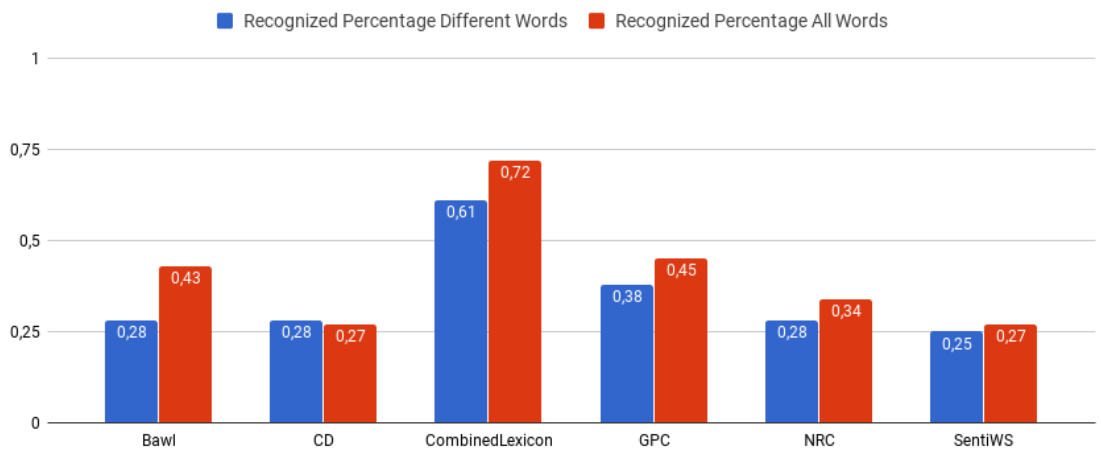


Abbildung 11: Vokabular-basierte Evaluation – Balkendiagramm – Lexikonvergleich mit Lemmatisierung

Man erkennt eine deutliche Steigerung der Erkennungsrate im Vergleich zu den nicht erweiterten Lexika, teilweise um 20-30%. Vergleich man die Lexika untereinander stellt man fest, dass Bawl, CD, NRC und SentiWS insgesamt ähnliche Ergebnisse bezüglich der Erkennung unterscheidbarer Wörter aufweisen. Das GPC hat die beste Erkennungsrate mit 38% sowohl für unterscheidbare Wörter, als auch alle Wörter zusammen. Auffällig ist auch hier wieder die geringere Steigerung der Erkennungsrate für SentiWS. Dies kann möglicherweise daran liegen, dass die DTA-Erweiterung auf Lexikon-Ebene auch zahlreiche Lemmas für jedes Wort miteinfügt (siehe Kapitel 5.3.3) und diese Flektionsformen bei SentiWS ohnehin schon vorliegen. Für das kombinierte Lexikon kann man festhalten, dass über die DTA-Erweiterung nun mehr als die Hälfte der unterscheidbaren Wörter als SBWs identifiziert wird, was nochmal die extreme Vergrößerung des Lexikon-Wortschatzes durch die DTA-Erweiterung deutlich macht. Es ist zweifelhaft ob tatsächlich 70% des Korpus aus SBWs besteht.

Neben dieser groben Zusammenfassung einiger Ergebnisse werden noch weitere Analyse-Ergebnisse genannt, die durch Einsicht der detaillierten Resultat-Dateien formuliert werden können:

- Lemmatisierung steigert die Erkennungsrate im Vergleich zur Ad-Hoc-Verwendung der Lexika.
- Bei DTA-erweiterten Lexika führt Lemmatisierung jedoch wenn überhaupt nur zu sehr geringen Verbesserungen.

- Es können keine größeren dramenspezifische Besonderheiten festgestellt werden. Prozentual betrachtet ist die Erkennungsrate bei kürzeren Dramen (z.B. Philotas) stets höher als bei längeren Dramen (z.B. Nathan der Weise)
- Bei Analyse des Korpus mit Stoppwörtern sind die Ergebnisse vor allem für die Erkennungsrate aller Wörter deutlich höher. Dies weist auf darauf hin, dass sich in einigen Lexika eine relevante Menge an Stoppwörtern befindet. Vor allem das GPC fällt auf, das in der Ad-Hoc-Verwendung 10% mehr Wörter erkennt als ohne Beachtung von Stoppwörtern.

Es werden keine weiteren Ergebnisse für die Vokabular-basierte Evaluation besprochen, da es sich um eine sekundäre Evaluationsmethode handelt. Die Resultate können im Detail dem Anhang entnommen werden.

6.4 Fazit

Über die Vokabular-basierte Evaluation konnten erste Unterschiede und Besonderheiten von Lexika und Verfahrensweisen identifiziert werden. Ferner konnten über die Erstellung und Analyse des Korpus-Wortschatzes Erkenntnisse über den Wortschatz und Sprachstil des Korpus erlangt werden. Das Framework kann zur ersten informellen Analyse in zukünftigen Projekten eingesetzt werden. Kann man aus unterschiedlichen Gründen keinen annotierten Evaluationskorpus erstellen, bietet die entwickelte Methodik zumindest eine oberflächliche Möglichkeit SA-Verfahren zu vergleichen und zu analysieren. Insgesamt kann über die Vokabular-basierte Evaluation jedoch nur ein grober und unscharfer Einblick in die Leistung einiger Lexika und Verfahren erlangt werden. Die Form der Evaluation ermöglicht keine aussagekräftigen Schlussfolgerungen über die tatsächliche Leistung eines Lexikons und genügt nicht dem im Projekt gestellten Anspruch einer systematischen Evaluation zur Identifikation eines optimalen Lexikon-basierten SA-Verfahrens. Aus diesem Grund wurde eine systematische Gold-Standard-Evaluation durchgeführt, die in den nachfolgenden Kapiteln beschrieben wird. Die Ergebnisse der Vokabular-basierten Evaluation werden als sekundär im Vergleich zu der Gold-Standard-Evaluation betrachtet.

7 Erstellung des Gold-Standard-Korpus

Die Erstellung 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 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, 2005) 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 Prob-

leme und Herausforderungen der Annotation literarischer Texte, analog zu Alm und Sproat (2005a) nun aber speziell für das deutschsprachige Drama des 18. Jahrhunderts, exploriert.

Die Erstellung des Gold-Standards 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.

7.1 Test-Korpus-Erstellung

7.1.1 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 Annotationen dieser Ebenen erforderen großes Expertenwissen. Im Rahmen der vorliegenden Arbeit bestand kein Zugriff auf zahlreiche domänenspezifische Experten. Ferner wäre eine diesbezügliche 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 (Alm & Sproat, 2005a) und die SA auf Sätzen ebenso als besonders herausfordernd betrachtet wird (Liu 2016, S. 70), 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).

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 4). In der Forschung werden beispielsweise 10% des Gesamtkorpus als Testkorpus gewählt (Alm et al., 2005). 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, 2005a). 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., 2014).

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 für Annotatoren als auch für die lexikonbasierte SA stellen kurze Sätze/Repliken größere Probleme dar (Alm & Sproat, 2005a; Mohammad, 2011; Nalisnick & Baird, 2013; Liu, 2016, S.10-11; S. 70). Annotatoren als auch die SA haben dann weniger Informationen zur Verfügung und die Entscheidung fällt in einem größeren 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 Kontexts 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 (2005a), 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 ein 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 7.1.2 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, das 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 7.1.3 besprochen.

7.1.2 Entwicklung

Zur Erstellung des Test-Korpus wurden 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 ausge-

baute 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.

7.1.3 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 folgen 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:

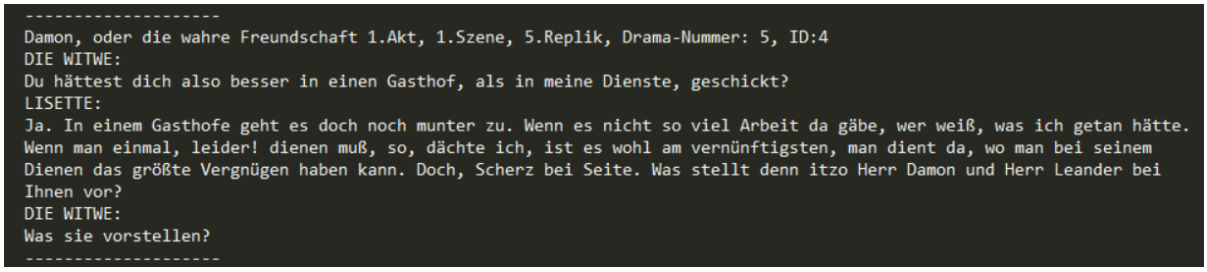


Abbildung 12: Ausschnitt Beispielreplik des Test-Korpus

Folgende Tabelle beschreibt die wesentlichen Statistiken des Test-Korpus:

Tabelle 16: Test-Korpus Statistiken

Zahl der Repliken	200
Positiv annotierte Repliken	61
Negativ annotierte Repliken	139

Kürzeste Replik	19
Längste Replik	306
Replikenlänge (Average)	50.675
Replikenlänge (Median)	38.0

Man erkennt, dass auf Basis der gewählten Kriterien der Testkorpus im Mittel deutlich längere Repliken aufweist, als der normale Gesamtkorpus. Repliken mit einer Länge von 19 sind dennoch recht kurz und bestehen meist nur aus 1-2 Sätzen. Für diese Fälle handelt es sich bei der implementierten und evaluierten SA um eine Form der Satz-basierten SA.

7.2 Test-Korpus-Annotation

7.2.1 Idee und Vorgehen

Zur Erhaltung eines Gold-Standards 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.

7.2.1.1 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 verlaufen 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 ist grundsätzlich die Angaben von positiv oder negativ für die Polarität denkbar. Zur Erhebung differenzierterer Informationen findet man jedoch Annotationsschemata mit mehr Auswahlmöglichkeiten, orientiert an Wiebe et al. (2005). Es werden einige Beispiele 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) und Saif et al. (2013) 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, das 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 (2005a) 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 2.4 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 unterschieden (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 auf Oberkategorien zurückgegriffen wird (Alm & Sproat, 2005a; Takala et al., 2014). 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-Erstellung 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:

Tabelle 17: Annotationsschema 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 des Annotationsverhaltens genauer betrachten, 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 (Alm & Sproat, 2005a; Alm et al., 2005; Volkova et al., 2010; Mohammad, 2011; Kakkonen und Kakkonen, 2011; Klinger et al., 2016). 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, 2005a). Das Annotationsschema wird durch folgende Tabelle illustriert. Teilnehmer geben lediglich an, ob eine Emotion vorliegt oder nicht:

Tabelle 18: Annotationsschema Emotionen

Zorn	Erwartung	Ekel	Angst	Freude	Traurigkeit	Überraschung	Vertrauen

7.2.1.2 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:

Der Freigeist 1.Akt, 5.Szene, 40.Replik, Drama-Nummer: 147, ID:14

JOHANN:

Je nu! so wird er das Geschäfte mit Ihnen so beiher treiben. Wir sind doch immer geklatscht.

ADRAST:

Du hast Recht. – – Ich möchte rasend werden, wenn ich an alle die Streiche gedenke, die mir ein ungerechtes Schicksal zu spielen nicht aufhört. – Doch wider wen murre ich? Wider ein taubes Ohngefähr? Wider einen blinden Zufall, der uns ohne Absicht und ohne Vorsatz schwer fällt? Ha! nichtswürdiges Leben! –

JOHANN:

O! lassen Sie mir das Leben ungeschimpft. So einer Kleinigkeit wegen sich mit ihm zu überwerfen, das wäre was Gescheutes!

Sehr negativ	Negativ	Neutral	Gemischt	Positiv	Sehr Positiv
	x				

Zorn	Erwartung	Ekel	Angst	Freude	Traurigkeit	Überraschung	Vertrauen
x							

Abbildung 13: Beispiel-Annotation

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 (2005a) lassen die Sätze aus Märchen von 4 Annotatoren in Zweier-Gruppen bewerten. Viele Studien nutzen jedoch eine ungerade Zahl von Anno-

tatoren, z.B. 3 (Momtazi, 2013; Takala et al., 2014; Eckle-Kohler, Kruge & Gurevych, 2015). Marchetti et al. (2014) 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, 2005a). 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, 2005a; 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 (2005a) 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.

7.2.1.3 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 7.2.2.6 werden die Daten des Fragebogens knapp ausgewertet.

7.2.2 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 letztes wird der in Kapitel 7.2.1.3 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.

7.2.2.1 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 systematischen Aufbereitung der Daten wurde ein Python-Programm entwickelt, das in Kapitel 7.2.2.2 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. Für eine Zusammenfassung aller Variablen wird auf die SPSS-Tabellen im 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 Annotatorenübergreifende 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.

7.2.2.2 Entwicklung und produzierte Datenstrukturen

Zur Berechnung differenzierter Übereinstimmungs-Metriken und zur holistischen Analyse dieser wurde ein Python-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.

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 Entwick-

lung 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)[5], 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 sind auch Maße für den Grad der Übereinstimmungen und werden deswegen in Kapitel 7.2.2.4 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.

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 Code verwiesen.

7.2.2.3 Sentiment-Verteilungen

Mittels der von der Annotation manuell aufbereiteten Daten und der Ausgabedaten, die im vorigen Abschnitt beschrieben, wurden, konnten die Sentiment-Verteilungen

[5] https://github.com/grrrr/krippendorff-alpha

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 (Alm & Sproat, 2005a; Volkova et al., 2010; Marchetti et al., 2014).

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.

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:

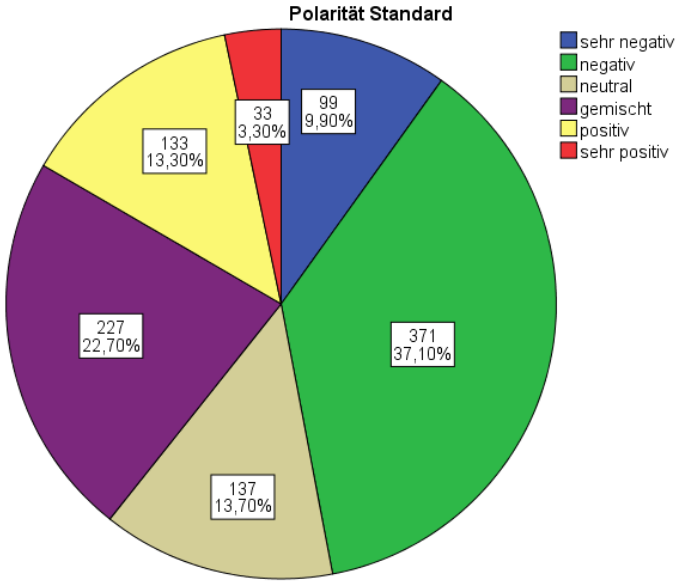


Abbildung 14: Sentiment-Verteilung Polarität Standard

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 (vier-

fach) betrachtet. Für diese Variable wurden die Gruppen positiv und sehr positiv zu einer Übergruppe positiv zusammengefasst. Selbiges gilt für die Negativitätsgruppen:

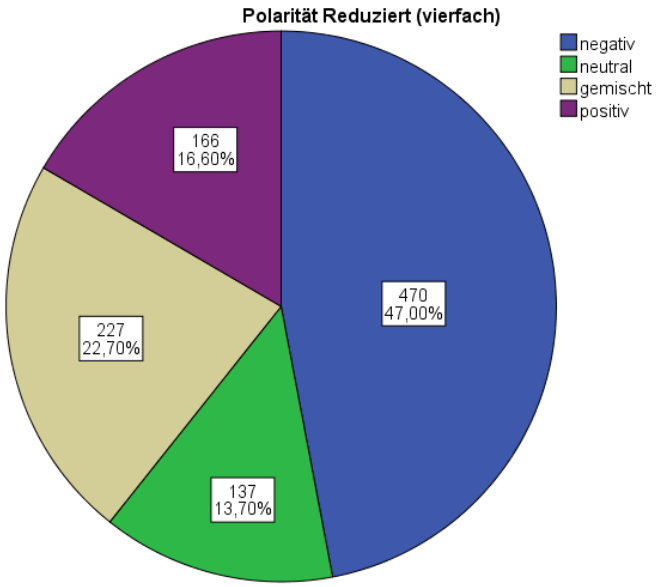


Abbildung 15: Sentiment-Verteilung Polarität Reduziert (vierfach)

Es wird eine besondere Ungleichverteilung deutlich. Die Annotatoren haben fast die Hälfte der Repliken 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 wenige 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:

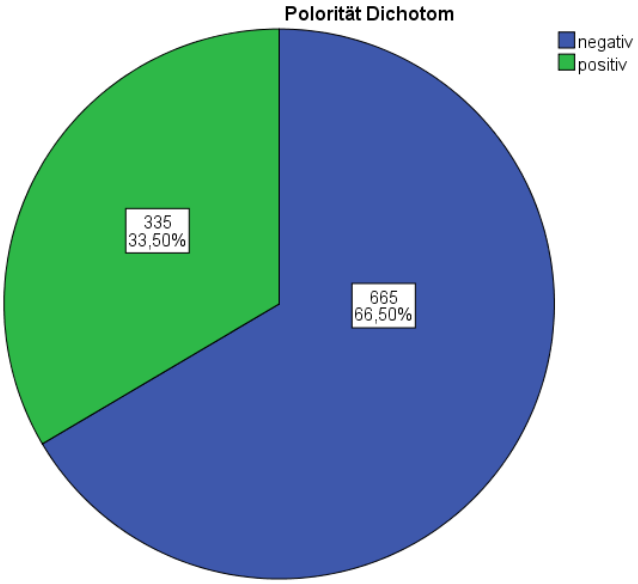


Abbildung 16: Sentiment-Verteilung Polarität Dichotom

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:

Tabelle 19: Kreuztabelle Polarität Reduziert (vierfach)*Polorität Dichotom

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

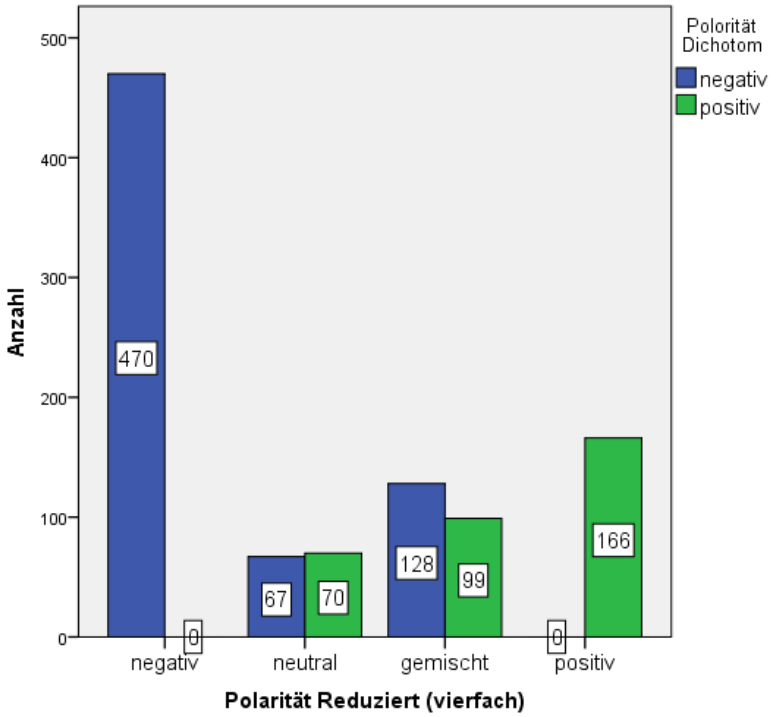


Abbildung 17: Balkendiagramm Polarität Reduziert (vierfach) * Polarität Dichotom

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:

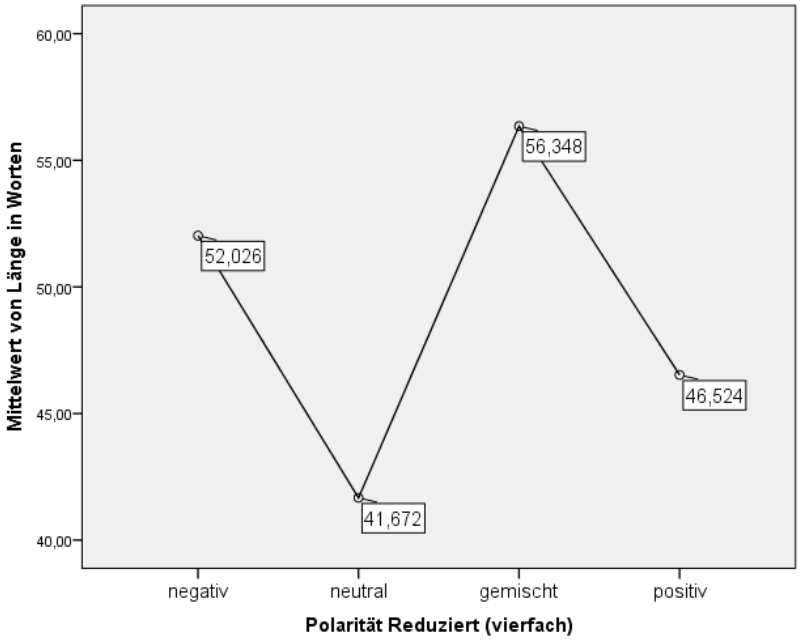


Abbildung 18: Mittelwert von Länge verteilt auf Polarität Reduziert (vierfach)

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.

Tabelle 20: Verteilung Emotionskategorien

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 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:

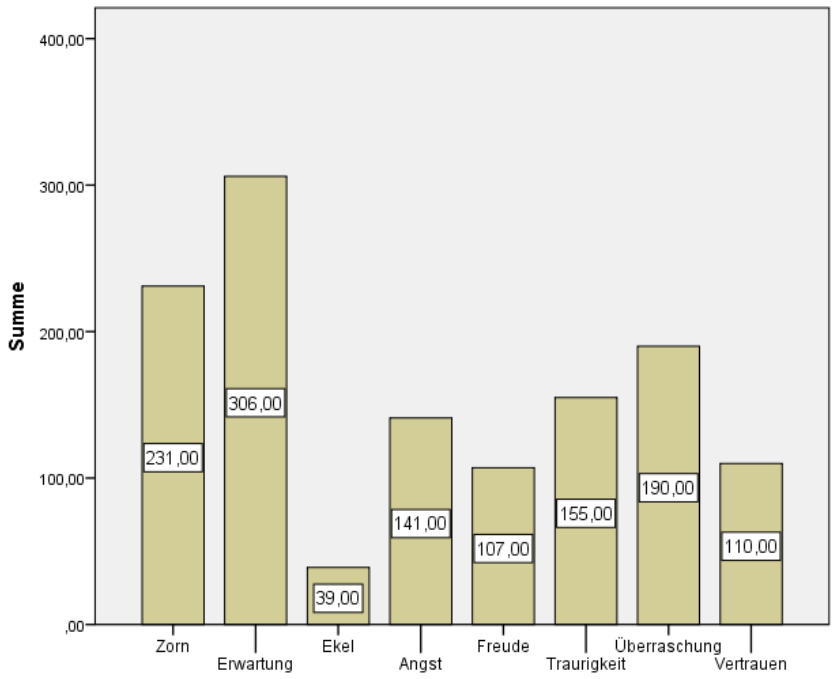


Abbildung 19: Balkendiagramm – Häufigkeitsverteilung Emotionskategorien

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:

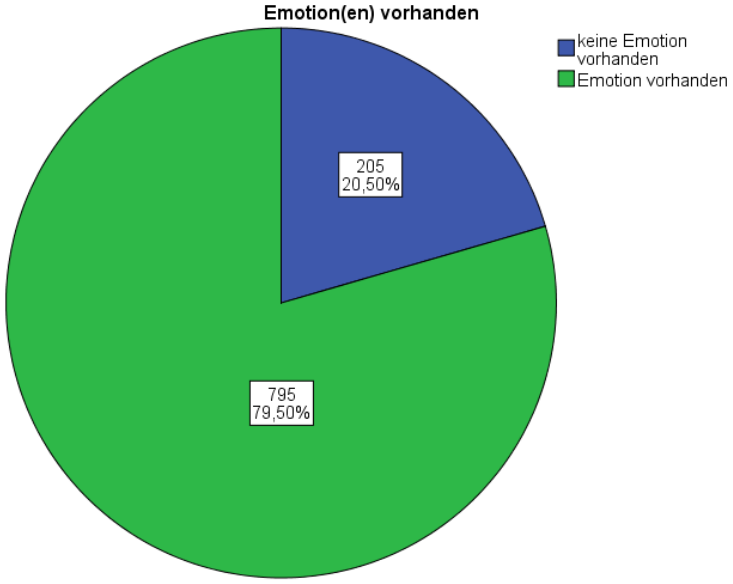


Abbildung 20: Häufigkeitsverteilung – Emotion vorhanden

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 wurden. Hier wird die Variable Polarität Dichotom gewählt (positiv, negativ):

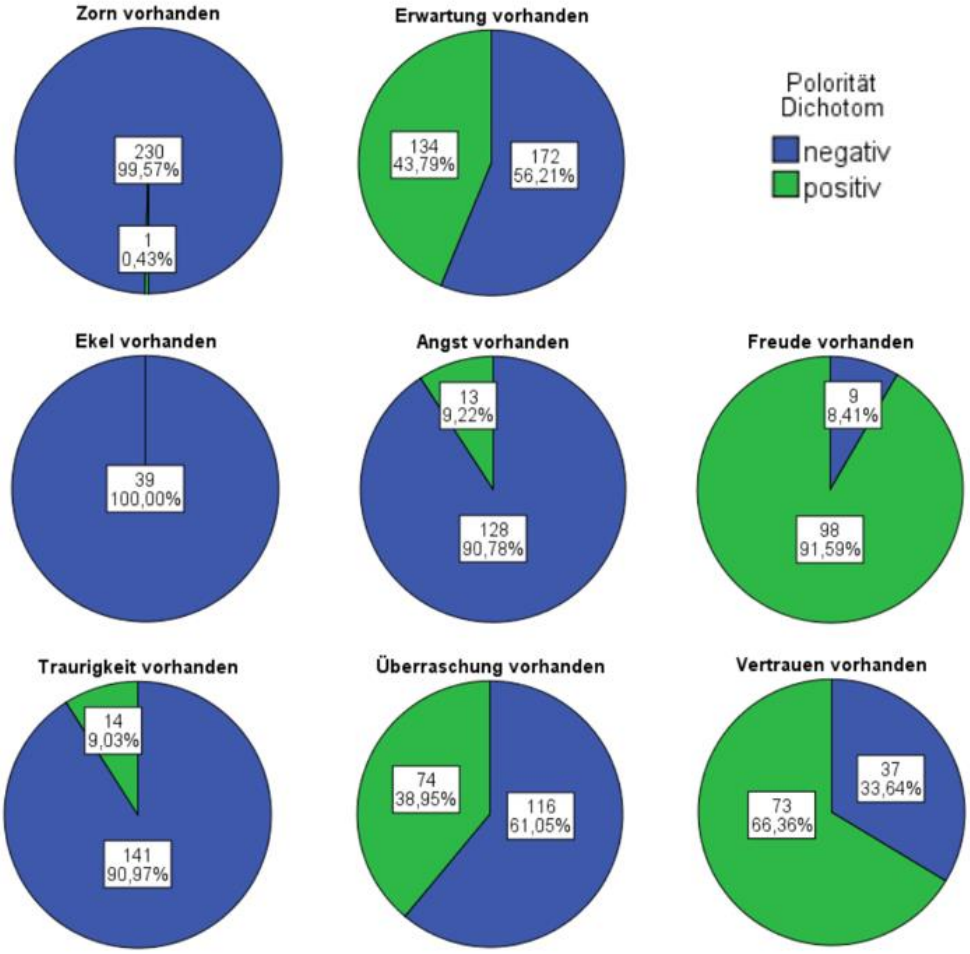


Abbildung 21: Häufigkeitsverteilungen – Emotionskategorien

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 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:

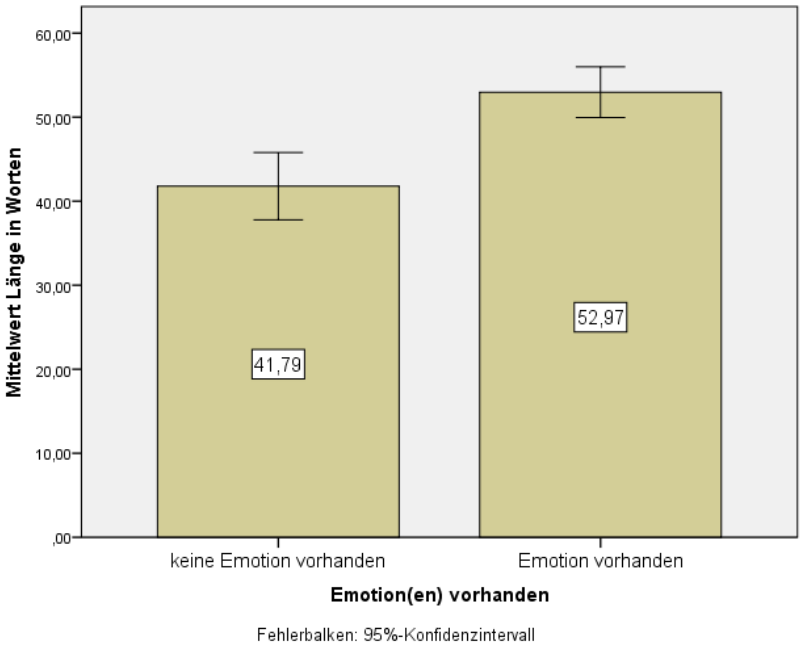


Abbildung 22: Mittelwerte von Länge für Emotion vorhanden

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.

7.2.2.4 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.

Alle Übereinstimmungsmaße wurden mittels den in Kapitel 7.2.2.2 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, das 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 (Momtazi, 2012; Takala et al., 2014). 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 (Krippendorff, 2011). 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 (Joyce, 2013; Antoine, Villaneau & Lefeuvre, 2014) 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).

$\alpha < 0$ = schlechte Übereinstimmung

$0 < \alpha < 0.2$ = schwache Übereinstimmung

$0.2 < \alpha < 0.4$ = mittelmäßige Übereinstimmung

$0.4 < \alpha < 0.6$ = moderate Übereinstimmung

$0.6 < \alpha < 0.8$ = substantielle Übereinstimmung

$0.8 < \alpha < 1$ = 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 7.2.2.3). Derartige Ungleichverteilungen führen bei K-Alpha und Fleiss' Kappa zu star-

ken Fehlkalkulationen (Feinstein & Cichetti, 1990; Gwet, 2011), je nach Ausmaß der Ungleichverteilung und je kleiner die Stichprobe der zu bewertenden Instanzen ist. So werden trotz hoher Übereinstimmung geringe Übereinstimmungsmaße berechnet. Dies wird von Feinstein und Cichetti (1990) als Paradox der Kappa-Statistiken bezeichnet. 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 um die Übereinstimmung deskriptiv zu beschreiben (Ku, Liang & Chen, 2006; Momtazi, 2012; Joyce, 2013; Bosco et al., 2014; Takala et al., 2014).

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, das die Übereinstimmung je nach gewählter Variable aufzeigt und den diesbezüglichen Trend verdeutlicht:

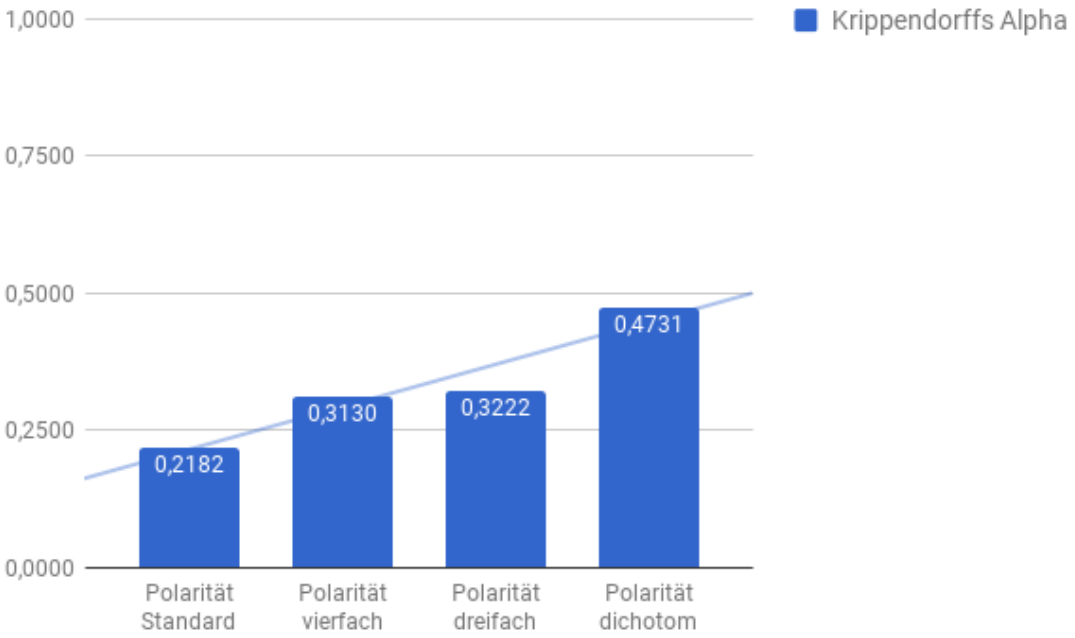


Abbildung 23: K-Alpha für Polaritätsvariablen

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:

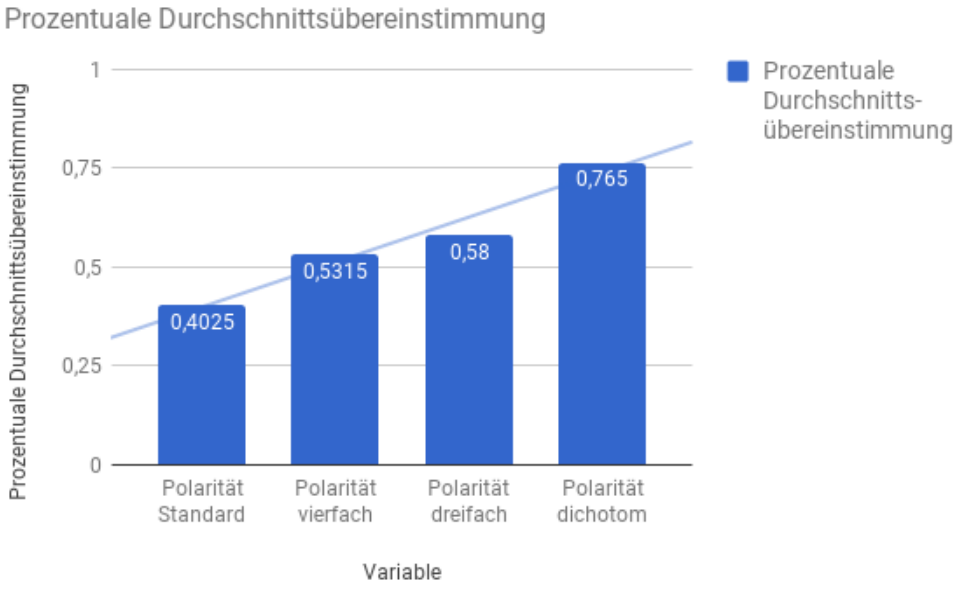


Abbildung 24: Prozentuale Durchschnittsübereinstimmung für Polaritätsvariablen

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. 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:

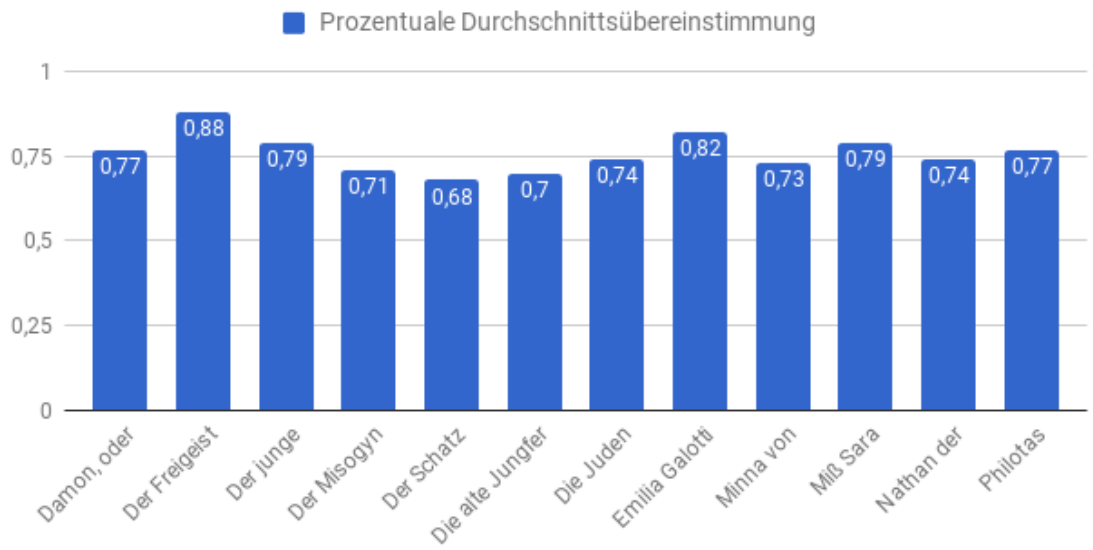


Abbildung 25: Prozentuale Durchschnittsübereinstimmung für Polarität Dichotom für alle Dramen

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 kein wichtiger dramenspezifischer Zusammenhang erkannt werden. 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 vom 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:

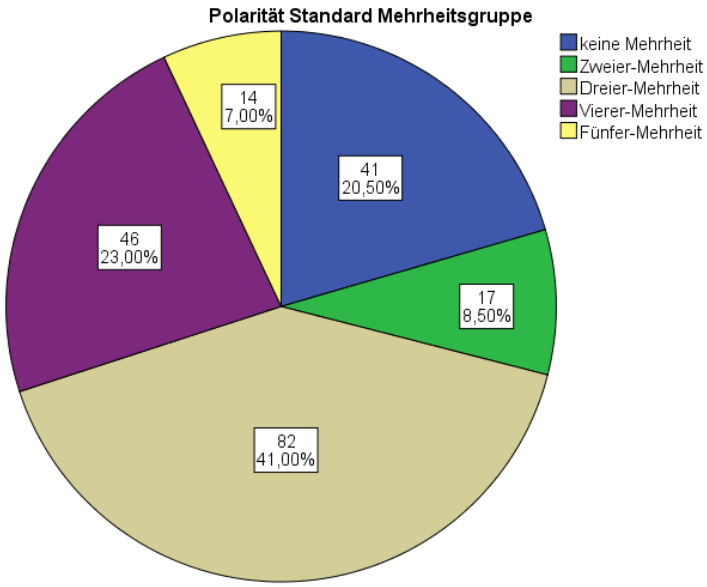


Abbildung 26: Häufigkeitsverteilung Mehrheitsgruppen Polarität Standard

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 Annotation 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:

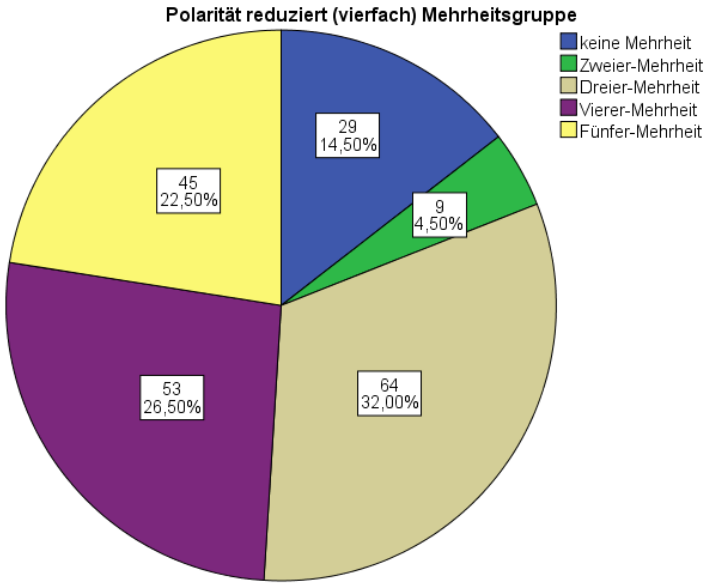


Abbildung 27: Häufigkeitsverteilung Mehrheitsgruppen Polarität reduziert (vierfach)

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:

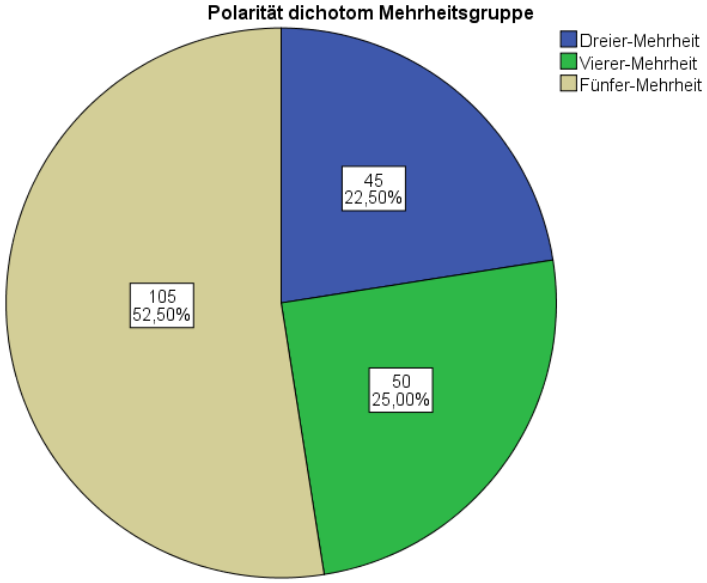


Abbildung 28: Häufigkeitsverteilung Mehrheitsgruppen Polarität Dichotom

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 vollkommener 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 Emotionsannotationen betrachtet. Aufgrund der deutlichen Ungleichverteilung dieser, sehr viel häufiger wird eine einzelne Emotion als nicht vorhanden annotiert (siehe Kapitel 7.2.2.3), liefern sie trotz hoher prozentualer Übereinstimmungen teilweise schwache Werte. Dies hängt mit bereits angesprochenen Problemen der Maße bei ungleichen Verteilungen zusammen. Aus diesem Grund werden hier nur die prozentualen Durchschnittsübereinst-

immungen angegeben. Die anderen Maße können jedoch im Anhang eingesehen werden:

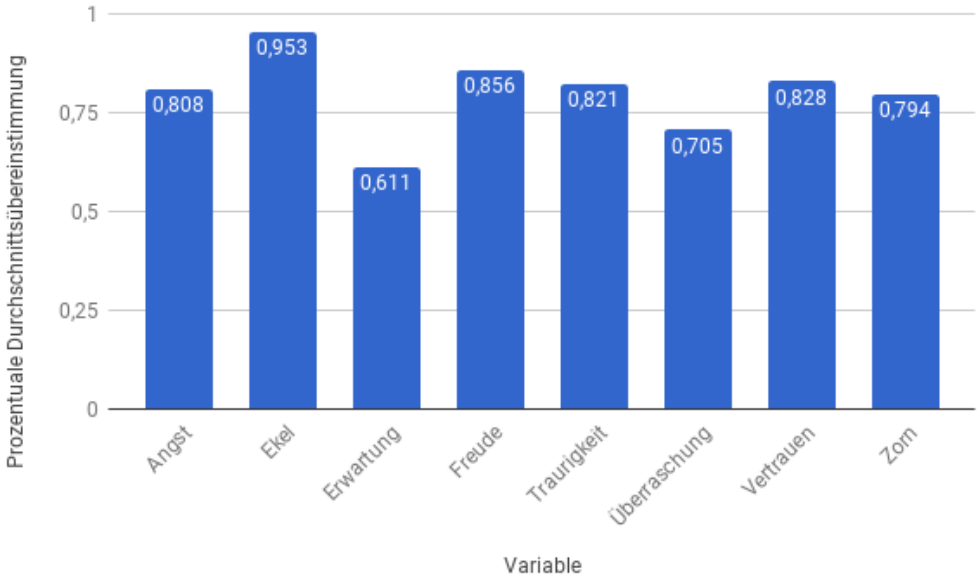


Abbildung 29: Prozentuale Durchschnittsübereinstimmung Emotionskategorien

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:

Tabelle 21: Mehrheitsverteilungen Emotionskategorien

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 Annotatoren eher übereinstimmen bei Fehlen einer Emotion 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 Mehrheitsgrup-

pen wie 3er- und 4er-Mehrheiten vorherrschen. Bezüglich dieser Emotionen sind sich Teilnehmer also insbesondere unsicher.

7.2.2.5 Mehrheitsannotationen

Abschließend werden nun noch die finalen Mehrheitsannotationen beschrieben. Darunter auch die finalen Annotationen für die Variable Polarität Dichotom, die in Kapitel 8 den Gold Standard und somit die Benchmark für die SA-Evaluation darstellt. Als finale Annotation einer 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 Verteilungen dieser finalen Sentiment-Auszeichnungen werden im Folgenden betrachtet. Sie ähneln grundsätzlich stark den allgemeinen Annotationsverteilungen (siehe Kapitel 7.2.2.3), 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.

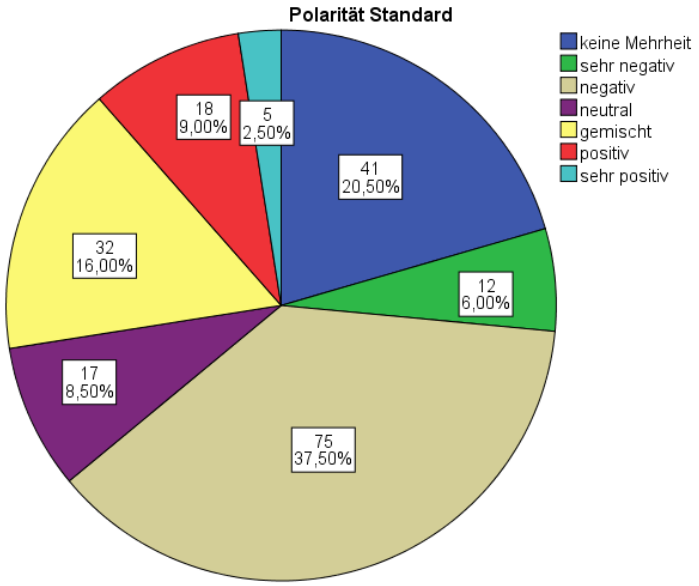


Abbildung 30: Häufigkeitsverteilung Mehrheitsentscheidungen Polarität Standard

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:

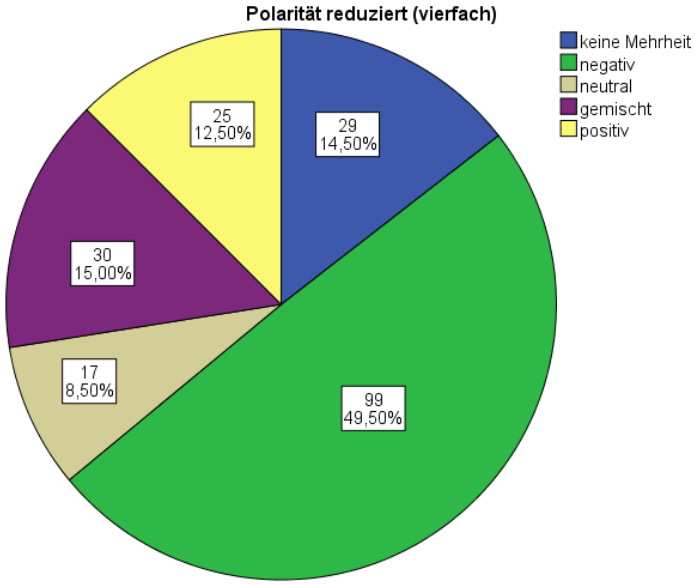


Abbildung 31: Häufigkeitsverteilung Mehrheitsentscheidungen Polarität reduziert (vierfach)

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:

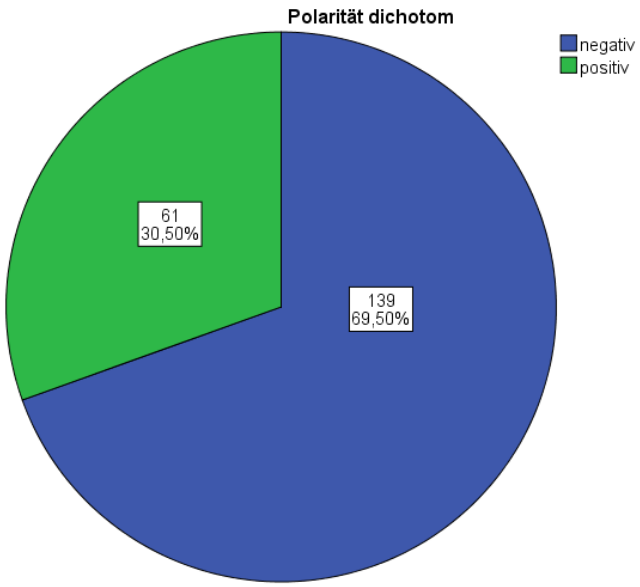


Abbildung 32: Häufigkeitsverteilung Mehrheitsentscheidungen Polarität Dichotom

Die finalen Annotationen für die Polarität Dichotom, basierend auf den Mehrheitsangaben, ergeben 61 positiv annotierte Repliken und 139 negative Repliken. In Kapitel 7.2.2.4 wurde bereits der Übereinstimmungsgrad bezüglich dieser Repliken besprochen, der insgesamt moderat ist (K-Alpha=0,47). 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.

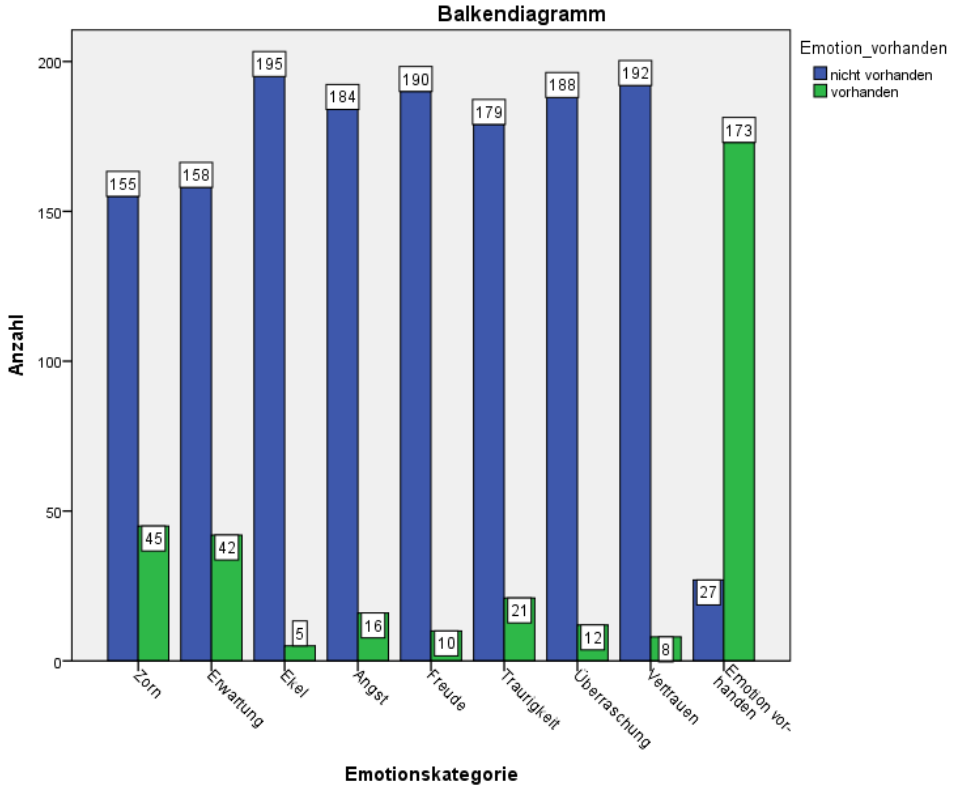


Abbildung 33: Häufigkeitsverteilung Mehrheitsentscheidungen Emotionskategorien

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.

7.2.2.6 Fragebogen-Auswertung

Im letzten Ergebnis-Abschnitt bezüglich der Test-Korpus-Annotation wird nun noch der Fragebogen 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.

Tabelle 22: Fragebogen – Schwierigkeit und Sicherheit – Statistiken

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:

Tabelle 23: Fragebogen – Zeit in Minuten – Statistiken

Deskriptive Statistiken

	N	Bereich	Minimum	Maximum	Mittelwert	Standardab-weichung	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 Probleme 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

7.3 Diskussion

Da es sich bei der Annotationsstudie, die in diesem Kapitel beschrieben wurde, um ein größeres relevantes Arbeitspaket der Forschungsagenda handelt, werden die Resultate nun separat diskutiert und interpretiert. Ähnlich wird für die Evaluationsstudie in Kapitel 8 verfahren. Die wichtigsten Erkenntnisse werden im Kontext der Forschung zusammengefasst und Grenzen und Probleme der verwendeten Methodik diskutiert. Auf Basis dieser werden Ideen für potentielle Anschlussstudien zum Thema Annotation auf literarischen Texten formuliert.

Es konnten detaillierte Einsichten in das Annotationsverhalten und die Konstitution des Korpus gewonnen werden. Es handelt sich nach Kenntnisstand des Autors um die erste Studie, die im Bereich literarischer Texte speziell Repliken bezüglich Sentiment annotieren lässt. Alm und Sproat (2005a) und Volkova et al. (2010) nutzen Sätze

aus Märchen, Marchetti et al. (2014) Sätze aus politischen Texten. Es war beim Studiendesign demnach nicht möglich auf bisherige Erfahrungen zurückzugreifen. Durch das gewählte Annotationsschema konnten differenzierte Erkenntnisse gesammelt werden. Bezüglich der Übereinstimmungsmaße verhalten sich die Resultate weitestgehend konform zur sonstigen verwandten Forschung. Für komplexere Polaritätsaufteilungen ist die Übereinstimmung sehr gering (K-Alpha = 0,22; Prozentuale Übereinstimmung = 40%), steigert sich aber zu einer durchschnittlichen Übereinstimmung bei einer simplen dichotomen Betrachtung (K-Alpha = 0,47; Prozentuale Übereinstimmung = 77%). Die Analysen von Mehrheitsverteilungen bestätigen dieses Bild. Bezüglich Emotionen sind sich Annotatoren eher einig, wenn diese nicht vorhanden sind, stimmen jedoch selten bezüglich des generellen Vorhandenseins überein. Besondere Probleme bereiten abstrakte Emotionen wie Überraschung und Erwartung. Die Ergebnisse zu den Übereinstimmungen sind insgesamt äquivalent zu den Resultaten von Alm und Sproat (2005a), Volkova et al. (2010) und Marchetti et al. (2014), die ebenso eher geringe Übereinstimmungen konstatieren, wobei man beachten muss, dass in diesen Studien andere Annotationsschemata verwendet wurden und Sätze betrachtet wurden. Innerhalb der sonstigen SA-Forschung liegen zwar auch Annotationsprojekte mit ähnlich geringen Übereinstimmungen vor, zum Beispiel für Tweets (Basile & Nissim, 2013), im Normalfall sind die Übereinstimmungen auf anderen Anwendungsgebieten jedoch höher für Kappa-Statistiken, z. B. bei 0,8 (Balahur & Steinberger, 2009). Auf Basis der bisherigen Forschung kann man also die These aufstellen, dass literarische Texte in höherem Maße von der individuellen Annotation abhängen. In der vorliegenden Arbeit wurden auch subjektive Angaben der Annotatoren zu Problemen erfasst. Hierbei wurde deutlich, dass die veraltete schwer verständliche Sprache, der mangelnde Kontext und häufige Polaritätswechsel während einer Replik große Probleme bereiten. In diesem Zusammenhang muss kritisch erwähnt werden, dass die Annotatoren keine besondere Ausbildung als Literaturwissenschaftler oder ähnliches haben. Dies ist nicht unüblich in verwandten Studien (Volkova et al, 2014). Man kann jedoch davon ausgehen, dass Experten, die besser mit der Sprache der Texte zurechtkommen und Kenntnisse über die Handlung der Dramen haben deutlich präzisere Annotationen durchführen können. Zukünftige Studie sollten die Akquise von Experten anstreben.

Bezüglich der Konstitution des Korpus kann man festhalten, dass vor allem gemischte (mehr noch als positive) aber auch neutrale Repliken einen relevanten Anteil darstellen. Diese Tatsache wurde bei der Umsetzung der SA und bei der kommenden Evaluation nicht weiter beachtet, sollte aber in zukünftigen Studien aufgegriffen werden. Denkbar ist eine Einführung dieser Klassen in die SA-Prädiktion oder eine Reduktion auf Satzebene um das Problem sowohl negativer als auch positiver Sentiments in einer Replik zu umgehen. Bezüglich Neutralität sei aber zu beachten, dass Alm und Sproat (2005a), Volkova et al. (2010) und Marchetti (2014) eine erhöhte Annotation von Sätzen als neutral feststellen konnten. Dieser Zusammenhang kann hier auf Repliken nicht konstatiert werden. Neutralität macht insgesamt einen eher kleineren Anteil aus (14% bei Polarität Reduziert (vierfach)).

Eine deutliche Besonderheit ist eine Ungleichverteilung bezüglich negativ annotierter Repliken. Gemäß Annotatoren wird die Mehrzahl der Repliken als negativ wahrgenommen, bei der dichotomen Aufteilung 67%, im finalen Korpus fast 70%. Dies entspricht den Ergebnissen von Alm und Sproat (2005a), wobei die Ungleichverteilung nicht in dem Ausmaß auftritt. Volkova et al. (2010) stellen eher das Gegenteil fest. Für das weitere Vorgehen in der SA aber auch in der Evaluation ist die Tatsache, dass das Korpus als eher negativ konnotiert wahrgenommen wird sehr relevant und zum Teil problematisch. Auf statistischer Ebene sind gewisse Evaluationsmaße nicht stabil gegenüber derartige Ungleichmäßigkeiten. Ferner werden aufgrund dieser Weise SA-Verfahren bevorzugt, die tendenziell eher negative Prädiktionen kalkulieren. In der Tat kann später bei Analysen auf höheren Ebenen als der Replik im Front-End festgestellt werden, dass das erhöhte Auftreten von Negativität ein festes Phänomen des Korpus ist. In Anbetracht bisheriger Forschung kann man vermuten, dass negative Polarität ein stabiler Bestandteil literarischer Texte ist. Es liegt an der Literaturwissenschaft eine Antwort auf dieses Befund zu finden. Mögliche Gründe sind möglicherweise, dass Konflikte und negative Ereignisse essentielle Bestandteile der meisten verwendeten Dramen sind und sich die Handlung vom ersten Akt an immer mehr hin zur Katastrophe oder dem plötzlichen glücklichen Ende entwickelt. Um Interesse zu wecken, müssen die Figuren negative Ereignisse durchleben, die sich dementsprechend in negativ konnotierten Dialogen und Repliken äußern. Die gemachten Behauptungen sind jedoch nur Vermutungen und können nicht explizit belegt werden. Weitere Projekte in

Zusammenarbeit mit der Literaturwissenschaft können diesen Aspekt jedoch im Detail untersuchen. Das Übergewicht von Negativität ist aber eine zentrale Erkenntnis, die bei der Interpretation von SA-Metriken stets beachtet werden muss.

Zuletzt sei noch kritisch die verhältnismäßig geringe Replikenzahl erwähnt, die in zukünftigen Projekten mit Experten oder mit interaktiven Tools, die die Annotation unterstützen, erhöht werden kann. Auch wird an dieser Stelle noch die Frage aufgeworfen ob Repliken tatsächlich die passende Ebene für die Annotation in der Dramenanalyse sind. Auf der untersten Ebene ließe sich noch der Satz betrachten auf höheren Ebenen die Szene oder komplexere Konzepte wie Sprecher und Beziehungen. Diese Ebenen können in weiteren Annotationsstudien exploriert werden.

8 SA-Evaluation

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

8.1 Vorgehen

Das Evaluationsvorgehen orientiert sich an herkömmlichen Vorgehen in der SA-Forschung (Turney, 2002; Pang & Lee, 2005; Goncalves et al, 2013; Zhou, Zhao & Shang, 2014; Agarwal et al., 2015; Ribeiro et al., 2016). 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 7.2.2.5 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 7.2.2.5). 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 hat. 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 2.3 und 5 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 5.5.2). Die Betrachtungsebene ist gemäß GS die Replik. Es kann eine Lexikonerweiterung mittels dem Tool von Jurish (2012) vorliegen (dtaExtended) oder nicht (noExtension). Wahlweise findet keine Lemmatisierung (noLemma), Lemmatisierung nur auf Text (textLemma) oder Lemmatisierung sowohl auf Text als auch 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 Kombinationen (z.B. caseInSensitive vs caseSensitive) 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 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 (Kaji & Kitsuregawa, 2008; Zhang et al., 2011; Ribeiro et al., 2016) eingesetzt. Zur Erklärungen dieser Maße wird sich an Goncalves et al. (2013) orientiert. Folgende Tabelle definiert einige Grundmaße zur Berechnung der letztendlichen Evaluations-Metriken:

Tabelle 24: Kreuztabelle – Prädiktionsmöglichkeiten

		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 werden die Repliken, die tatsächlich positiv sind, aber vom SA-Verfahren eine negative Bewertung erhalten, angegeben. 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

durchführt, 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 Evaluationsmaßes Genauigkeit mit Recall, Precision und F-Wert ist in der vorliegenden Studie zwingend notwendig, da die Polaritäten sehr ungleich verteilt sind. Ein SA-Verfahren, das 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 Verfahrens 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.

8.2 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 Projekten 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. Innerhalb von Test_Corpus_Evaluation wird der Gold-Standard als Liste initialisiert sowie der als Pickle-Datei abgespeicherte Test-Korpus 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, das 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 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.

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.

8.3 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 8.3.4 werden dann die Ergebnisse auch vergleichend zusammengefasst, zentrale Aussagen formuliert und die besten SA-Verfahren vorgestellt.

8.3.1 Datenaufbereitung

Die tsv-Tabellen, die über das in Kapitel 8.2 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:

	A	B	C	D	E	F	G	H	I	J
1	CombinationType	DTAExtension	Lemmatizer	LemmatizationType	Stopwords	CaseSensitivity	accuracy	recallPositive	precisionPositive	F-MeasurePositive
2	noExtension_toker	noExtension	tokens	noLemma	noStopwordList	caseInSensitive	0,38	0,524590163	0,2519685039	0,3404255319
3	noExtension_toker	noExtension	tokens	noLemma	noStopwordList	caseSensitive	0,405	0,4754098361	0,25	0,3276836158
4	noExtension_toker	noExtension	tokens	noLemma	standardList	caseInSensitive	0,38	0,524590163	0,2519685039	0,3404255319
5	noExtension_toker	noExtension	tokens	noLemma	standardList	caseSensitive	0,405	0,4754098361	0,25	0,3276836158
6	noExtension_toker	noExtension	tokens	noLemma	enhancedList	caseInSensitive	0,41	0,4262295082	0,2385321101	0,3058823529
7	noExtension_toker	noExtension	tokens	noLemma	enhancedList	caseSensitive	0,4	0,393442623	0,2242990654	0,2857142857
8	noExtension_toker	noExtension	tokens	noLemma	enhancedFiltered	caseInSensitive	0,405	0,4918032787	0,2542372881	0,3351955307
9	noExtension_toker	noExtension	tokens	noLemma	enhancedFiltered	caseSensitive	0,405	0,4754098361	0,25	0,3276836158
10	noExtension_treeta	noExtension	treetagger	textLemma	noStopwordList	caseInSensitive	0,415	0,6721311475	0,2971014493	0,4120603015
11	noExtension_treeta	noExtension	treetagger	textLemma	noStopwordList	caseSensitive	0,425	0,6393442623	0,2954545455	0,4041450777
12	noExtension_treeta	noExtension	treetagger	textLemma	standardList	caseInSensitive	0,415	0,6721311475	0,2971014493	0,4120603015
13	noExtension_treeta	noExtension	treetagger	textLemma	standardList	caseSensitive	0,425	0,6393442623	0,2954545455	0,4041450777
14	noExtension_treeta	noExtension	treetagger	textLemma	enhancedList	caseInSensitive	0,445	0,3770491803	0,2395833333	0,2929936306
15	noExtension_treeta	noExtension	treetagger	textLemma	enhancedList	caseSensitive	0,445	0,3770491803	0,2395833333	0,2929936306
16	noExtension_treeta	noExtension	treetagger	textLemma	enhancedFiltered	caseInSensitive	0,42	0,6229508197	0,2900763359	0,3958333333

Abbildung 34: Ausschnitt Ergebnistabellen Evaluationsstudie

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.

8.3.2 Benchmark

Zur Interpretation der Erkennungsleistung der SA-Verfahren muss eine Benchmark bestimmt werden, also eine Leistung, die mindestens erreicht bzw. übertroffen werden muss, um von einer ausreichend guten Leistung zu sprechen. In der SA-Forschung gibt es verschiedene Ansätze diese Benchmark zu definieren. Zentrale Benchmark-Metrik ist meist die accuracy (Zhang et al., 2011; Hu et al., 2013).

Als Mindestleistung wird häufig orientiert an üblichen Evaluationsverfahren im Maschinellen Lernen die random baseline gewählt (Davidov et al., 2010a; Elsner, 2012), welche die Leistung angibt, die ein System erreichen würde, das alle Klassen von Untersuchungseinheiten (im SA meist die Polaritäten positiv und negativ) zufällig auswählt. Die grundsätzliche Idee ist also, dass ein System zumindest besser sein sollte als der Zufall. Die random baseline wird als Summe der quadrierten Anteile der Klassen am Gesamtkorpus berechnet (Davidov et al., 2010a). Die Anteile der Klassen sind die Wahrscheinlichkeiten, dass eine bestimmte Klasse im Korpus vorkommt. Bei gleichverteilten Klassen ist die random baseline demnach immer 0,5. Wie in Kapitel 7.2.2.3 beschrieben wurde, liegt im vorliegenden Korpus keine Gleichverteilung vor. Die Wahrscheinlichkeit einer negativen Replik im Korpus liegt bei 0,695, einer positiven bei lediglich 0,205. Dies ergibt aufgrund obiger Rechnung eine random baseline von 0,525. Bei derartigen Ungleichverteilungen kann als strengerer Vergleichsmaßstab die majority baseline gewählt werden (Nakov et al., 2013). Die majority baseline ist die Leistung, die ein System produzieren würde, das konsistent die häufigste Klasse zuweisen würde. Dieses System würde im vorliegenden Fall eine accuracy von 0,695 (Anteil an negativen Repliken am Korpus) erzeugen. Aufgrund der starken Ungleichverteilung hinsichtlich negativer Repliken ist die majority baseline sehr hoch. Ogneva (2010) und Mozetic et al. (2016) schlagen Übereinstimmungen zwischen menschlichen Annotatoren als Benchmark vor. Die Idee dabei ist, dass ein maschinelles System nur so gut sein kann, wie sich Menschen bei der Beurteilung desselben Untersuchungsgegenstandes übereinstimmend sicher sind. Die Übereinstimmung von Menschen wird damit als menschliche Klassifikationsleistung definiert und als anzustrebender Maßstab des eigenen Systems. Mozetic et al. (2016) verwenden diesbezüglich beispielsweise Maße wie Krippendorffs Alpha um verschiedene Evaluationsergebnisse vergleichend zu analysieren. Auch im Bereich literarischer Texte verwenden Marchetti et al. (2014) das Maß

Fleiss-Kappa zwischen Annotatoren als Vergleichswert. Ogneva (2010) weist auf die Übereinstimmung in Prozent hin. Diese Idee, diese Maße als Benchmark zu verwenden wird hier aufgegriffen. Es konnte auch bereits gezeigt werden, dass die Übereinstimmung für den vorliegenden Anwendungsfall geringer ist als in anderen Bereichen. K-Alpha für Polarität Dichotom ergibt 0,47 und liegt damit unter der random baseline, weswegen dieses Maß als Benchmark verworfen wird. Die prozentuale Übereinstimmung hingegen beträgt 0,765 und ist damit von allen bisher genannten Benchmark-Werten der höchste. Es ist naheliegend bei vergleichbaren Anwendungsgebieten und Aufgabenstellungen den eigenen SA-Ansatz mit der Leistung anderer Projekte auf dem gleichen Anwendungsgebiet zu vergleichen (Nakov et al., 2013). Es liegen auch Studien vor, die bei der Analyse eines fortgeschrittenen neuen SA-Verfahrens ein primitiveres (z.B. die ad-hoc-Nutzung eines SA-Lexikons) Verfahren als Vergleichs-Benchmark verwenden (Hu et al., 2013; Agarwal et al., 2015). Da es nach Kenntnisstand des Autors keine ähnlichen Studien gibt, die systematisch die SA auf literarischen Texten untersuchen, ist ein Vergleich nicht möglich, wobei man natürlich die einzelnen hier eingesetzten Methoden untereinander verglichen werden können. Lediglich Marchetti et al. (2014) führen Evaluationen durch, jedoch nur auf einem sehr kleinen Korpus von historischen Texten und auf Satzebene. Sie erhalten dabei ein Durchschnittsergebnis von 43% Genauigkeit über den Einsatz von SA-Lexika.

In Anbetracht des Mangels an vergleichbaren Studien und da zum erste Mal die SA-Leistung auf einem Dramenkorpus untersucht wird, hat man sich für folgende differenzierte Auswahl an Benchmark-Werten entschieden. Als Mindestleistung wird die random baseline von 0,525 gewählt. Als Kennwerte für besonders leistungsstarke SA-Verfahren wird auch stets die majority baseline von 0,695 und die prozentuale Übereinstimmung von 0,765 betrachtet. Ferner werden zur genauen Leistungsanalyse neben der accuracy auch stets die anderen polaritätsspezifischen Leistungmaße betrachtet.

8.3.3 Evaluationsergebnisse pro Lexikon und Polaritäts-Metrik

Die Ergebnisse werden nun pro Lexikon und Metrik zusammengefasst präsentiert. Es wird die durchschnittliche Leistung sowie die beste und schwächste Leistung tabellarisch zusammengefasst. Dann werden die besonders leistungsstarken Verfahren aufgelistet und diskutiert. Wie bereits angemerkt kann hier nur ein Ausschnitt aller Ergebnisse gezeigt werden, da 800 verschiedene Herangehensweisen mittels etwa 10 Pefor-

manzmaßen analysiert werden. Dennoch werden oft Aussagen über Gesamtanalysen gemacht, weswegen zur genauen Einsicht auf den Anhang verwiesen wird.

Die Ergebnisse werden zumeist mittels zweier Tabellen pro Polaritätsmetrik beschrieben. Zunächst werden übersichtsweise die Durchschnittsergebnisse, Leistungs-Minima und Maxima für polaritySentiWS präsentiert und danach die besten Ergebnisse besprochen. Unter dem Durchschnitt wird die Summe alle Werte eines Maßes bezüglich aller 80 Herangehensweisen geteilt durch 80 verstanden. Der Wert repräsentiert die allgemeine Durchschnittsleistung der Polaritätsmetrik bezogen auf alle Herangehensweisen. Das Minimum die schlechteste Kombinatorik an Methoden und das Maximum die beste. Die einzelnen Methoden die die jeweiligen Werte ergeben können unterschiedlich sind. Konkrete Methoden werden dann in der zweiten nach Leistung sortierten Tabelle beschrieben und diskutiert. Die Ergebnisse werden auf 3 Stellen gerundet angegeben. Die nachweislich besseren Lexika und Polaritäts-Metriken werden ausführlicher behandelt und analysiert als die klar weniger leistungsstarken.

8.3.3.1 Sentiment Wortschatz – SentiWS

Für SentiWS liegen zwei Polaritäts-Metriken vor. Die Kalkulation der Polarität über die angegeben Polaritätsgewichte (polaritySentiWS) sowie über das simple Term-Zähl-Verfahren, bei dem die gewichteten SentiWS-Angaben dichotomisiert wurden (polaritySentiWSDichotom) (siehe Kapitel 5.5.2).

Zunächst die Übersicht über die Leistung der Herangehensweisen über die Metrik polaritySentiWS:

Tabelle 25: Übersicht Evaluationsmaße polaritySentiWS

Evaluationsmaß	Minimum	Durchschnitt	Maximum
accuracy	0,53	0,587	0,705
F-Measure Positive	0,328	0,433	0,528
F-Measure Negative	0,610	0,6725	0,797

Es fällt auf, dass die Genauigkeit aller Methoden für polaritySentiWS selbst bei der schlechtesten über der random baseline liegt, bei der besten Methodik mit einer Erkennung von 0,705 über der majority baseline. Bezüglich der F-Werte kann man übersichtsweise ausmachen, dass die Erkennung negativer Repliken deutlich besser ausfällt

als für positive. Selbst die beste SA-Methode zur Erkennung von positiven Repliken erzeugt lediglich einen F-Wert von 0,53, während negativere Repliken selbst bei schlechteren Verfahren einen guten Gesamtwert erbringen (Min: 0,61). Zur besseren Analyse werden die Ergebnisse der 6 besten Verfahren und des schlechtesten angegeben und kurz besprochen.

Tabelle 26: Ausschnitt Evaluationsmaße Verfahrenskombinationen polaritySentiWS

DTAExtension	Lemmatizer	Lemmatization Type	Stopwords	Case Sensitivity	Accuracy	F-Measure Positive	F-Measure Negative
dtaExtended	treetagger	textLemma	noStopwordList	caseInSensitive	0,705	0,4587	0,797
dtaExtended	treetagger	textLemma	noStopwordList	caseSensitive	0,695	0,429	0,791
dtaExtended	textblob	textLemma	noStopwordList	caseInSensitive	0,675	0,3925	0,778
dtaExtended	tokens	noLemma	noStopwordList	caseInSensitive	0,67	0,410	0,770
dtaExtended	textblob	textLemma	noStopwordList	caseSensitive	0,67	0,365	0,777
dtaExtended	textblob	bothLemma	noStopwordList	caseSensitive	**0,67**	**0,528**	**0,746**
…	…	…	…	…	…	…	…
noExtension	tokens	noLemma	enhancedList	caseSensitive	0,53	0,328	0,638

Es wird insgesamt der gewinnbringende Nutzen der meisten Verarbeitungsverfahren deutlich. Die schlechteste Leistung wird quasi bei einer ad-hoc-Verwendung des Lexikons, bei der lediglich eine Stoppwortliste eingesetzt wird, konstatiert. Sowohl Lemmatisierung als auch die Lexikonerweiterung steigern die Genauigkeit. Die beste accuracy wird mittels Lexikonerweiterung, Lemmatisierung der Textgrundlage mittels treetagger, keiner Stoppwortliste und keiner Beachtung der Groß- und Kleinschreibung erreicht. Dadurch werden insgesamt in dem Fall 144 der 200 Repliken korrekt erkannt. Letztgenanntes Attribut (Groß- und Kleinschreibung) ändert jedoch meistens nur marginal die Leistung. Die Wahl des Lemmatisierers ändert ebenso nur geringfügig die Leistung, entscheidend ist die Lexikonerweiterung. Das beste Verfahren ohne diese hat lediglich eine accuracy von 0,575.

Es fällt insgesamt auf, dass die besten Verfahren keine Stoppwortlisten einsetzen. Aus diesem Grund wurden die Ergebnisse noch im Detail analysiert hinsichtlich eines problematischen Einflusses von Stoppwörtern auf die SA, der fälschlicherweise zu positiver Leistung führt. Tatsächlich resultiert die Lexikonerweiterung und Lemmatisierung darin, dass Stoppwörter wie er und ihr als SBWs markiert werden. Es kann jedoch kein besonderer und vor allem konsistenter Einfluss dieser festgestellt werden. Zunächst sind diese Stoppwörter etwas seltener in den Repliken als andere Stoppwörter (z.B. Artikel siehe Kapitel 8.3.3.5 und die Polaritätsmetrik von GPC). Des Weiteren sind diese sowohl positiv als auch negativ annotiert und die Gewichtung sehr schwach. Die Werte gleichen sich damit quasi aus und sind zu gering in der Gewichtung um besonderen Einfluss zu erhalten. Ferner fällt keine extreme Ungleichverteilung bezüglich der Klassifikation auf. Zwar zeigt sich ein Übergewicht negativer Vorhersagen, jedoch gilt dieses Übergewicht unabhängig vom Einsatz von Stoppwortlisten. Aus diesen Gründen wurden die Verfahren trotz problematischer Stoppworteinflüsse als leistungsstarke SA-Verfahren beibehalten.

Ein besonderer Fokus muss noch auf Methode 6 der obigen Tabelle gelegt werden. Über die beidseitige Lemmatisierung und Lexikonerweiterung wird „sein“ als starkes positives SBW markiert und bei der Kalkulation genutzt. Dies ist grundsätzlich ein Fehler, resultiert aber in überdurchschnittlich guten Leistungen, besonders bezüglich der Vorhersage positiver Repliken, wobei die Vorhersage negativer auch noch überdurchschnittlich gut bleibt. Insgesamt hat die Methode den besten durchschnittlichen F-Wert (0,637) und erkennt 97 negative Repliken korrekt und 37 positive. Im Gegensatz dazu erkennt die beste Methodik gemäß accuracy 116 negative Repliken korrekt und 25 positive. Trotz der insgesamt schwächeren Leistung wirkt das Stoppwort „sein“ für Methode 6 wie eine Art positiver Gewichtungsfaktor, die dadurch besonders starke und häufige negative Wörter (die aber tatsächlich da sind) benötigt, um Repliken als negativ zu klassifizieren. Das Resultat führt dann insgesamt zu einer polaritätsbezogen gleichmäßigeren und guten Erkennungsleistung. Dennoch ist Methode 1 insgesamt bezogen auf accuracy am besten, obschon sie nur 36% der positiven Repliken korrekt erkennt.

Es wird nun die Leistung der SentiWS-Metrik nach simpler Term-Zähl-Methodik betrachtet (polaritySentiWSDichotom):

Tabelle 27: Übersicht Evaluationsmaße polaritySentiWSDichotom

Evaluationsmaß	Minimum	Durchschnitt	Maximum
accuracy	0,395	0,433	0,505
F-Measure Positive	0,331	0,411	0,472
F-Measure Negative	0,331	0,449	0,591

Es fällt sofort auf, dass die Ergebnisse im Schnitt und auch bezüglich der besten Verfahren deutlich schlechter ausfallen als bei der gewichteten Metrik von SentiWS. Das beste Verfahren ist schlechter als die random baseline. Die F-Werte sind im Schnitt ähnlich, analog zu polaritySentiWS ist die Metrik leistungsstärker bei negativen Repliken. Zur Vollständigkeit werden auch hier die besten 5 und das schlechteste Verfahren präsentiert, zur genaueren Einsicht aber auf den Anhang verwiesen.

Tabelle 28: Ausschnitt Evaluationsmaße Verfahrenskombinationen polaritySentiWSDichotom

DTAExtension	Lemmatizer	Lemmatization Type	Stopwords	Case Sensitivity	Accuracy	F-Measure Positive	F-Measure Negative
dtaExtended	tokens	noLemma	noStopwordList	caseInSensitive	0,505	0,414	0,571
dtaExtended	textblob	textLemma	noStopwordList	caseInSensitive	0,505	0,407	0,575
dtaExtended	tokens	noLemma	noStopwordList	caseSensitive	0,5	0,404	0,568
dtaExtended	treetagger	textLemma	noStopwordList	caseInSensitive	0,5	0,418	0,561
dtaExtended	textblob	textLemma	noStopwordList	caseSensitive	0,5	0,397	0,572
…	…	…	…	…	…	…	…
noExtension	treetagger	bothLemma	noStopwordList	caseInSensitive	0,395	0,447	0,331

Die grundsätzlichen Erkenntnisse, die bereits bei polaritySentiWS gemacht werden konnten, dass die Lexikonerweiterung den stärksten Leistungsboost erbringt, werden hier bestätigt. Bezüglich der Lemmatisierung kann kein eindeutiges Leistungsschema erkannt werden. Die Lexikonerweiterung mit linguistischen und historischen Varianten ist ausreichend, die Lemmatisierer beeinflussen die Leistung sonst nicht mehr be-

sonders. Zwei Verfahren sind etwa gleich gut, die Version mit Lexikonerweiterung, ohne Stoppwörter und ohne Groß- und Kleinschreibung, einmal lemmatisiert im Text über Textblob und einmal ohne jegliche Lemmatisierung.

Beide Verfahren erkennen genau 101 Repliken exakt, also etwa die Hälfte. Es fällt auf, dass die Leistung bei der positiven Prädiktion deutlich schlechter ist. Dies war bereits bei polaritySentiWS der Fall, kann dort aber auf einen schlechteren Recall zurückgeführt werden, wohingegen hier die Precision für positive Repliken sehr schwach ist, d.h. es werden übermäßig viele Repliken als positiv klassifiziert. Methode 1 beispielsweise produziert 73 false positives mit einer Precision Positive von 0,32. Diese positive Überbewertung steht im deutlichen Widerspruch zur Polaritäts-Metrik polaritySentiWS. Der Widerspruch, dass für polaritySentiWS zwar mehr positive Wörter erkannt werden, diese aber weniger starke Gewichte haben bzw. die negativen Wörter sind seltener, haben aber größere Gewichte.

Für den Stoppworteinfluss gelten ähnliche Aussagen wie bereits oben. Die besten Leistungen identifiziert man bei Verfahren ohne Stoppwortlisten. Der Einfluss ist jedoch nicht so deutlich. Das nächstbeste Verfahren ohne Stoppwörter erzielt bereits die achtbeste Leistung mit einer accuracy von 0,465. Auch hier kann kein deutlicher Einfluss von Vorhandensein von Stoppwortlisten oder nicht festgestellt werden, es gilt generell, dass das Problem zu viele false positives sind. Fälschlicherweise negative und positive Stoppwörter gleichen sich insgesamt wieder aus. Aus diesem Grund werden die Methoden hier nicht weiter herausgefiltert. Die spezielle Besonderheit der Stoppwörter muss jedoch trotz des marginalen Einflusses bei der Gesamtanalyse beachtet werden.

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

Auch das BAWL-R besteht aus einer Metrik mit Polaritätsstärken emotion und einer Term-Zähl-Metrik polarityBawlDichotom.

Als erstes wird die gewichtete Metrik emotion betrachtet:

Tabelle 29: Übersicht Evaluationsmaße emotion

Evaluationsmaß	Minimum	Durchschnitt	Maximum
accuracy	0,365	0,425	0,485
F-Measure Positive	0,429	0,480	0,507
F-Measure Negative	0,180	0,350	0,460

Die Übersichtsergebnisse belegen eine schwache Gesamtleistung. Die beste Methodik ist unter der random baseline. Die F-Werte weisen im Schnitt auf sehr große Probleme bezüglich der Vorhersage negativer Repliken hin. Tatsächlich weisen die Gesamtergebnisse (siehe Anhang) im Schnitt einen sehr hohen positiven Recall mit schwacher Precision auf. Das Gegenteil ist für die negative Klassifikation der Fall. Das heißt es findet eine extrem starke Überklassifikation von positiven Repliken statt. Die fünf besten und das schlechteste Verfahren verdeutlichen dieses Problem:

Tabelle 30: Ausschnitt Evaluationsmaße Verfahrenskombinationen emotion

DTAExtension	Lemmatizer	Lemmatization Type	Stopwords	Case Sensitivity	Accuracy	F-Measure Positive	F-Measure Negative
dtaExtended	treetagger	bothLemma	noStopwordList	caseInSensitive	0,485	0,507	0,460
dtaExtended	treetagger	bothLemma	noStopwordList	caseSensitive	0,48	0,5	0,458
dtaExtended	textblob	textLemma	enhancedList	caseInSensitive	0,465	0,472	0,456
dtaExtended	textblob	textLemma	enhancedList	caseSensitive	0,465	0,472	0,456
dtaExtended	textblob	bothLemma	enhancedList	caseInSensitive	0,465	0,478	0,451
…	…	…	…	…	…	…	…
dtaExtended	treetagger	textLemma	noStopwordList	caseSensitive	0,365	0,481	0,180

Auch hier kann man insgesamt einen positiven Einfluss der Lexikonerweiterung erkennen. Ferner kann man für emotion bis auf die ersten beiden Verfahren einen positiven Einfluss von Stoppwortlisten identifizieren. Nach den zwei besten Methoden kommt die nächstbeste Methode ohne Stoppwörter erst wieder auf Platz 44. Dies liegt

daran, dass mit der enhancedList insgesamt erfolgreich positive häufige Wörter des Korpus entfernt werden konnten.

Dennoch ist das Hauptproblem, die häufige Auszeichnung von Repliken als positiv grundsätzlich für jede Methode erkennbar. Für die beste Methode bezüglich accuracy kann man 53 true positives, 95 false positives, 44 true negatives und 8 false negatives identifizieren. Das BAWL-R identifiziert die Mehrzahl der Repliken als positiv. Die Precision für die negativen Repliken ist sehr gut, doch aufgrund dessen, dass deutlich mehr negative Repliken im Korpus enthalten sind, entstehen insgesamt schlechte Evaluationswerte.

Das gleiche Schema wird auch für die dichotome Metrik deutlich polarityBawlDichotom:

Tabelle 31: Übersicht Evaluationsmaße polarityBawlDichotom

Evaluationsmaß	Minimum	Durchschnitt	Maximum
accuracy	0,325	0,367	0,405
F-Measure Positive	0,394	0,446	0,473
F-Measure Negative	0,106	0,258	0,370

Wieder erkennt man, dass die Gesamtleistung diesmal sehr deutlich auch für die besten Methodenkombination unter der random baseline liegt. Das Problem der schlechten Klassifikation negativer Repliken bleibt bestehen. Auch die Auflistung der 5 besten und der schlechtesten Leistung belegt dies.

Tabelle 32: Ausschnitt Evaluationsmaße Verfahrenskombinationen polarityBawlDichotom

DTAExtension	Lemmatizer	Lemmatization Type	Stopwords	Case Sensitivity	Accuracy	F-Measure Positive	F-Measure Negative
dtaExtended	treetagger	textLemma	enhancedList	caseInSensitive	0,405	0,436	0,370
dtaExtended	treetagger	textLemma	enhancedList	caseSensitive	0,405	0,436	0,370
noExtension	tokens	noLemma	enhancedFilteredList	caseInSensitive	0,4	0,464	0,318
noExtension	treetagger	bothLemma	enhancedFilteredList	caseSensitive	0,4	0,473	0,302
dtaExtended	textblob	textLemma	enhancedList	caseInSensitive	0,4	0,449	0,340
…	…	…	…	…	…	…	…
noExtension	treetagger	bothLemma	noStopwordList	caseInSensitive	0,325	0,444	0,140

Der eben genannte Befund wird bestätigt. Die einzelnen Methodenkombinationen sind dabei unbedeutend. Die beste Methode erkennt auch hier lediglich 81 Repliken korrekt mit 104 false positives. Die schlechteste Methode produziert 128 false positives. Insgesamt scheitert die SA also an der großen Überklassifikation von Positivität.

Um diese Feststellung der Überklassifikation näher zu analysieren wurden informell einige Repliken über die Detailevaluations-Dateien (siehe Kapitel 8.2) analysiert. Es konnte jedoch tatsächlich keine Besonderheit identifiziert werden. Es werden sowohl passende negative als auch positive SBWs gefunden, jedoch meist mehr positive SBWs, die auch gleichzeitig höhere Gewichte haben. Es konnte jedoch kein besonderer Einfluss von bestimmten Wörtern oder ein anderer Fehler festgestellt werden. Das BAWL-R produziert für das gewählte Dramenkorpus gemäß korrekter Kalkulation übermäßig positive Klassen und ist aufgrund des höheren Anteils an Negativität im Korpus ungeeignet für das vorliegende Projekt.

8.3.3.3 NRC Word-Emotion Association Lexikon (NRC)

Für das NRC gibt es lediglich eine Polaritätsmetrik, da keine Gewichte vorliegen, sondern nur Zugehörigkeiten zu den Klassen positiv und negativ. Die Metrik polarityNrc basiert demnach auf simpler Term-Zähl-Methodik:

Tabelle 33: Übersicht Evaluationsmaße polarityNrc

Evaluationsmaß	Minimum	Durchschnitt	Maximum
accuracy	0,34	0,421	0,525
F-Measure Positive	0,288	0,393	0,465
F-Measure Negative	0,310	0,441	0,588

Man erkennt, dass die accuracy von polarityNrc im Schnitt unter der random baseline liegt und im besten Fall exakt auf dieser (0,525). Bezüglich der beiden Polaritäten kann man mittels der Durchschnitts- und Maximum-Werte erkennen, dass die Probleme etwas größer bei der Klassifikation von positiven Repliken sind. Der Unterschied ist jedoch nicht übermäßig groß.

Tabelle 34: Ausschnitt Evaluationsmaße Verfahrenskombinationen polarityNrc

DTAExtension	Lemmatizer	Lemmatization Type	Stopwords	Case Sensitivity	Accuracy	F-Measure Positive	F-Measure Negative
dtaExtended	treetagger	bothLemma	noStopwordList	caseInSensitive	0,525	0,437	0,588
dtaExtended	treetagger	bothLemma	enhancedList	caseInSensitive	0,5	0,404	0,568
dtaExtended	textblob	textLemma	noStopwordList	caseSensitive	0,495	0,416	0,555
dtaExtended	textblob	textLemma	noStopwordList	caseInSensitive	0,49	0,406	0,552
dtaExtended	treetagger	bothLemma	enhancedList	caseSensitive	0,49	0,385	0,564
…	…	…	…	…	…	…	…
noExtension	textblob	textLemma	enhancedFilteredList	caseSensitive	0,34	0,352	0,326

Die Detailanalyse zeigt, dass sowohl Lexikonerweiterung als auch Lemmatisierung, sowohl auf Text als auch auf Lexikon-Ebene die Genauigkeit aber auch die polaritätsbezogenen F-Werte steigert. Vor allem die Lexikonerweiterung stellt die größte Verbesserung dar, die beste Methodik ohne diese hat eine Genauigkeit von 0,445. Sowohl die Stoppwortlisten, als auch die Groß- und Kleinschreibung hat nur einen marginalen Einfluss. Dieser Befund legt nahe, dass das NRC wenige Stoppwörter oder andere häufige Wörter des Korpus enthält.

Die beste Methode gemäß accuracy verwendet eine Lexikonerweiterung mit beidseitiger Lemmatisierung ohne Stoppwörter und ohne auf Groß- und Kleinschreibung zu achten. Es werden 105 Repliken korrekt erkannt. Es liegen jedoch 71 false positives und 24 false negatives vor. Ähnlich zum BAWL-R (siehe Kapitel 8.3.3.2), jedoch in deutlich geringerem Ausmaß neigt die NRC-Metrik zur vermehrten und falschen Klassifikation von Repliken als positiv. Dies gilt für alle Herangehensweisen. Es liegt meist ein sehr hoher Recall mit geringer Precision für positive Vorhersagen vor, was wiederum bei den negativen Vorhersagen genau umgekehrt ist, da wenige Repliken als negativ bewertet werden und wenn dann meist korrekt. Auch hier können durch eine informelle Analyse der exakten Kalkulation der SBWs per Replik keine weiteren Erkenntnisse gewonnen werden. Die Kalkulation verläuft korrekt, es werden deutlich mehr positive Wörter gefunden. Auch das NRC kann aufgrund seiner Übergewichtung mit positiven SBWs im Korpus als nicht gewinnbringend für die Polaritätsbestimmung im vorliegenden Projekt betrachtet werden. Als einziges genutztes Lexikon mit komplexeren Emotionsangaben liefert es dennoch einen wichtigen Mehrwert.

8.3.3.4 Clematide-Dictionary (CD)

Für das CD liegen wieder zwei Polaritäts-Metriken vor. Die Metrik polarityCd nutzt die angegebenen Polaritätsstärken zur Kalkulation, polarityCdDichotom weist jedem positiven und negativen Wort zur Aufrechnung generell den Wert 1 zu. Im Fall des CD sind die Unterschiede zwischen diesen beiden Metriken gering, da die einzigen Polaritätsstärken für polarityCd die Ausprägungen 0,7 und 1 annehmen können.

Es wird zunächst das gewichtete Maß polarityCd und seine Leistung betrachtet:

Tabelle 35: Übersicht Evaluationsmaße polarityCd

Evaluationsmaß	Minimum	Durchschnitt	Maximum
accuracy	0,43	0,503	0,605
F-Measure Positive	0,343	0,462	0,553
F-Measure Negative	0,402	0,534	0,677

Man kann feststellen, dass die CD-Methoden im Schnitt eine Genauigkeit von 0,5 erreichen, also die Hälfte der Repliken korrekt erkennen. Tatsächlich zeigt das Maximum aber auch eine Erkennungsrate die mit 0,605 über der random baseline aber unter der

majority baseline liegt. Die F-Werte lassen wieder erkennen, dass die Hauptprobleme bei der Klassifikation von positiven Repliken liegen, jedoch insgesamt nicht so problematisch wie bei anderen Lexika sind. Man stellt wieder einen hohen Recall und eine geringe Präzision bei positiven Repliken fest, sowie einen durchschnittlichen Recall mit einer hohen Precision bei negativen Repliken. Dies wird nach Präsentation der

Tabelle 36: Ausschnitt Evaluationsmaße Verfahrenskombinationen polarityCd

DTAExtension	Lemmatizer	Lemmatization Type	Stopwords	Case Sensitivity	Accuracy	F-Measure Positive	F-Measure Negative
dtaExtended	treetagger	bothLemma	enhancedFilteredList	caseSensitive	0,605	0,553	0,645
dtaExtended	treetagger	bothLemma	enhancedFilteredList	caseInSensitive	0,6	0,550	0,639
dtaExtended	textblob	bothLemma	enhancedList	caseInSensitive	0,595	0,514	0,652
dtaExtended	treetagger	bothLemma	enhancedList	caseInSensitive	0,59	0,438	0,677
dtaExtended	treetagger	bothLemma	enhancedList	caseSensitive	0,59	0,438	0,677
…	…	…	…	…	…	…	…
noExtension	tokens	noLemma	standardList	caseInSensitive	0,43	0,393	0,462

Insgesamt wird für das CD ein grundsätzlich positiver Einfluss aller Methoden deutlich. Die besten Leistungen werden durch Lexikonerweiterung und durch eine Lemmatisierung auf Lexikon- und Text-Ebene über den treetagger-Lemmatisierer erreicht. Auch der Einsatz der erweiterten Listen enhancedList und enhancedFilteredList ist gewinnbringend, zu häufige Wörter des Korpus und Stoppwörter werden erfolgreich herausgefiltert und die Genauigkeit und Präzision des Verfahrens gesteigert. Methoden ohne Stoppwortliste oder mit lediglich der Standard-Liste sind meist schlechter (siehe Gesamttabelle im Anhang). Dies zeigt auch, dass vor allem die Filterung von Korpus-spezifischen häufigen Wörtern gewinnbringen ist für die Performanz des CD. Die Groß- und Kleinschreibung hat wieder insgesamt einen vernachlässigbaren Einfluss.

Die beste Methode verwendet eine Lexikonerweiterung mit beidseitiger treetagger-Lemmatisierung, der enhancedFilteredList als Stoppwortliste und der Beachtung von Groß- und Kleinschreibung im letzten Abgleichschritt. Auf diese Weise werden

etwas mehr als 60% der Repliken korrekt erkannt (121 Repliken). Das grundsätzliche Problem von CD und eben auch hier der besten Methodik ist die hohe Zahl an false positives, die zu einer schlechten Precision für positive Repliken führt. Das Ausmaß ist jedoch nicht so deutlich wie bei anderen Lexika (BAWL-R, NRC). Es werden 49 true positives (insgesamt 61 positive Repliken), 67 false positives, 72 true negatives (insgesamt 139 negative Repliken) und lediglich 12 false negatives von der CD-Methodik ausgegeben. Auch hier kann eine Detail-Analyse keinen besonderen Fehler oder ein überhäufiges Wort identifizieren, das die oben gezeigten besten CD-Verfahren bei der Kalkulation aufweisen.

Die dichotome Version der CD-Polarität unterscheidet sich per Definition kaum von ihrer Version mit Polaritätsstärken. Dies zeigt sich auch an ähnlichen Evaluationsergebnissen:

Tabelle 37: Übersicht Evaluationsmaße polarityCdDichotom

Evaluationsmaß	Minimum	Durchschnitt	Maximum
accuracy	0,39	0,448	0,53
F-Measure Positive	0,317	0,416	0,509
F-Measure Negative	0,306	0,471	0,617

Die Ergebnisse sind vom Verlauf her sehr ähnlich. Das Hauptproblem liegt wieder an einer schlechteren Erkennungsleistung von positiven Repliken, die von einer geringen Precision mit erhöhtem Recall geprägt ist. Es fällt jedoch klar über die accuracy auf, dass diese deutlich schlechter ist als bei der gewichteten Version der CD-Polarität. Sie liegt mit einem Wert von 0,53 nur knapp über der random baseline und weit unter der besten Methode für die Metrik cdPolarity. Zur Vollständigkeit werden jedoch auch hier die wichtigsten Verfahren tabellarisch zusammengefasst:

Tabelle 38: Ausschnitt Evaluationsmaße Verfahrenskombinationen polarityCDDichotom

DTAExtension	Lemmatizer	Lemmatization Type	Stopwords	Case Sensitivity	Accuracy	F-Measure Positive	F-Measure Negative
dtaExtended	treetagger	bothLemma	enhancedList	caseSensitive	0,53	0,389	0,617
dtaExtended	treetagger	bothLemma	enhancedList	caseInSensitive	0,525	0,387	0,612
dtaExtended	treetagger	bothLemma	enhancedFilteredList	caseSensitive	0,52	0,483	0,551
dtaExtended	treetagger	bothLemma	enhancedFilteredList	caseInSensitive	0,515	0,481	0,544
dtaExtended	textblob	bothLemma	enhancedList	caseInSensitive	0,5	0,438	0,549
…	…	…	…	…	…	…	…
dtaExtended	textblob	bothLemma	noStopwordList	caseSensitive	0,39	0,455	0,306

Die einzelnen Methoden wirken fast identisch wie bei cdPolarity, weswegen auf obigen Abschnitt zur Interpretation der Daten verwiesen wird. Der einzige Unterschied ist, dass die Leistung für cdPolarityDichotom grundsätzlich schlechter ist bezogen auf alle Evaluationsmaße. Das beste hier vorliegende Verfahren ist das gleich bei cdPolarity. Es werden 106 von 200 Repliken korrekt erkannt. Der große Unterschied zwischen cdPositive und cdPolarityDichotom wird über die F-Werte der positiven Klasse deutlich. Dieser ist sehr viel schlechter. Der Grund hierfür ist ein deutlich schlechterer Recall, es werden viel weniger Repliken korrekt als positiv erkannt, nämlich lediglich 30 bei 63 fälschlich so bewerteten (zum Vergleich: bei positiveCd 49 zu 67). Die Leistung bezüglich negativer Repliken ist dabei fast gleich geblieben.

Zusammenfassend kann man sagen, dass ähnlich zu SentiWS die gewichtete Version der Polaritäts-Metrik des CD die bessere Metrik zur SA auf dem Korpus ist. Durch eine Optimierung mittels verschiedener Methoden kann auch eine zufriedenstellende Erkennungsgenauigkeit von 0,605 erreicht werden, die über der Zufallsgrenze liegt.

8.3.3.5 German Polarity Clues (GPC)

Für das German Polarity Clues gibt es lediglich eine Polaritäts-Metrik nach Term-Zähl-Verfahren ohne Polaritätsstärken: polarityGpc.

Bezüglich dieses Lexikons und dieser Metrik liegt eine Besonderheit vor, die zur Anpassung der Daten geführt hat. Aus diesem Grund werden nun zunächst die grundsätzlich besten Verfahren präsentiert und die Probleme dieser besprochen:

Tabelle 39: Ausschnitt Evaluationsmaße Verfahrenskombinationen polarityGpc

DTAExtension	Lemmatizer	Lemmatization Type	Stopwords	Case Sensitivity	Accuracy	F-Measure Positive	F-Measure Negative
dtaExtended	textblob	textLemma	noStopwordList	caseInSensitive	0,7	0,230	0,813
dtaExtended	textblob	textLemma	noStopwordList	caseSensitive	0,7	0,230	0,813
noExtension	treetagger	bothLemma	noStopwordList	caseInSensitive	0,695	0,246	0,808
dtaExtended	treetagger	bothLemma	noStopwordList	caseInSensitive	0,695	0,246	0,808
noExtension	treetagger	bothLemma	noStopwordList	caseSensitive	0,69	0,243	0,805
…	…	…	…	…	…	…	…
dtaExtended	treetagger	bothLemma	enhancedList	caseSensitive	0,495	0,348	0,587

In der Tat kann man bei alleiniger Betrachtung der accuracy verhältnismäßig gute Ergebnisse konstatieren. Die besten Verfahren liegen alle über der random baseline und der majority baseline. Das hier angegebene Verfahren erkennt 140 Repliken korrekt. Bei genauer Betrachtung kann man jedoch feststellen, dass die F-Werte für die positiven Repliken auffällig schlecht und die F-Werte für die negativen besonders gut sind. Eine Detailanalyse über die vollständigen Tabellen (siehe Anhang) identifiziert man sehr hohe Recall-Werte (>0,9) für die negative Prädiktion und sehr geringe Recall-Werte für die positive (<0,2). Um das Problem des GPC zu verdeutlichen wird hier für das gemäß accuracy beste Verfahren die Kreuztabelle zur genauen Erkennungsleistung angegeben:

Tabelle 40: Prädiktionstabelle polarityGpc

		Vorhergesagte Erwartungen (SA-Verfahren)	
		Negativ	Positiv
Tatsächliche Beobachtungen (Gold Standard)	Negativ	131 (True Negatives)	8 (False Positives)
	Positiv	52 (False Negatives)	9 (True Positives)

Die Überklassifikation von negativen Repliken ist deutlich. Lediglich 17 Repliken überhaupt werden als positiv markiert. Dies gilt für alle der leistungsstärkeren Verfahren. Das System weist also fast ausschließlich Repliken der Klasse negativ zu. Ein derartiges System, das grundsätzlich negative Prädiktionen vornimmt, erzeugt aufgrund des übermäßigen Anteils negativer Repliken am gesamten Test-Korpus immer überdurchschnittlich gute Erkennungsraten über der random baseline und nah an der majority baseline. Letztere ist nämlich dadurch definiert, die accuracy anzugeben, die man erhält, wenn man stets die häufigste Klasse zuweist. Die vorliegende Metrik verhält sich fast exakt auf genau diese Art und Weise und generiert auf diese Weise Resultate, die oberflächlich betrachtet zufriedenstellend sind.

Bei genauerer Betrachtung wird deutlich, dass dies an dem übermäßigen Einfluss von Stoppwörtern liegt. Die 20 Verfahren, die keine Stoppwortlisten nutzen, sind gleichzeitig die 20 besten gemäß Genauigkeit. Sonstige Optionen wie Lexikonerweiterung und Lemmatisierung haben darauf fast keinen Einfluss. Lediglich der Einsatz des textblob-Lemmatisierers ohne Lexikonerweiterung gleicht die starke Überklassifikation ohne Stoppwortlistenverwendung etwas aus. Der starke Einfluss wird deutlich, wenn man Ergebnisse mit Stoppwortlisten betrachtet. Hier kann man dann entgegen der bisherigen Ergebnisse eine leichte Überklassifikation bezüglich Positivität feststellen. Eine detaillierte Analyse auf Wortebene zeigt, dass einige eindeutig nicht-sentiment tragende Wörter wie Artikel (der, dem) oder das Verb „sein“ als negative SBWs identifiziert werden. Diese entstehen vor allem durch Probleme in der Lemmatisierung und durch Lexikonerweiterung. Sie sind zum Teil (z.B. dem) auch tatsächlich bereits in der Roh-Form des Lexikons enthalten. Alle gefundenen Stoppwörter sind dabei negativ annotiert. Da es sich hier auch nicht um eine gewichtete Metrik handelt, sondern um Term-Zähl-Verfahren ist der Einfluss auf die Gesamt-Klassifikation noch größer. Derartige Interaktionen mit nicht sentiment-tragenden Stoppwörtern können auch bei anderen Lexika festgestellt werden, haben dort jedoch aus unterschiedlichen ausgleichenden Gründen nicht denselben starken Einfluss wie hier. Da die Ergebnisproduktion aufgrund der Stoppwörter, obschon äußerlich gute Ergebnisse erlangt werden, grundsätzlich fehlerbelastet ist, hat man sich dazu entschieden, die vorliegenden SA-Verfahren ohne Stoppwortlisten herauszufiltern und des Weiteren nur jene mit Stoppwortlisten zu betrachten.

Filtert man diese fehlerhaften Herangehensweisen heraus, weist polarityGpc folgende Gesamtergebnisse auf:

Tabelle 41: Übersicht Evaluationsmaße polarityGpc

Evaluationsmaß	Minimum	Durchschnitt	Maximum
accuracy	0,495	0,538	0,605
F-Measure Positive	0,3378378378	0,451	0,526
F-Measure Negative	0,5377358491	0,599	0,660

Trotz der Filterung erreicht man mit der Metrik polarityGpc im Schnitt noch immer Erkennungsraten über der random baseline, im besten Fall mit einem Wert von 0,6. Die majority baseline wird somit nicht erreicht. Es wird im Schnitt deutlich, dass mit Stoppwortlisten tatsächlich die Erkennungsleistung, gemessen am F-Wert erkennbar besser bei positiven Repliken ist als ohne Stoppwortlisten. Folgende sind die besten sowie das schlechteste Verfahren:

Tabelle 42: Ausschnitt Evaluationsmaße Verfahrenskombinationen polarityGpc (gefiltert)

DTAExtension	Lemmatizer	Lemmatization Type	Stopwords	Case Sensitivity	Accuracy	F-Measure Positive	F-Measure Negative
dtaExtended	textblob	textLemma	enhancedFilteredList	caseSensitive	0,605	0,526	0,660
dtaExtended	textblob	textLemma	enhancedFilteredList	caseInSensitive	0,6	0,523	0,655
dtaExtended	tokens	noLemma	enhancedFilteredList	caseInSensitive	0,59	0,506	0,649
dtaExtended	tokens	noLemma	enhancedFilteredList	caseSensitive	0,59	0,506	0,649
dtaExtended	textblob	textLemma	enhancedList	caseInSensitive	0,59	0,487	0,658
…	…	…	…	…	…	…	…
noExtension	treetagger	bothLemma	enhancedList	caseSensitive	0,495	0,348	0,587

Man erkennt in der Gesamttabelle konsistent wieder eine Verbesserung der Leistung durch Lexikonerweiterung. Der Zusammenhang ist jedoch nicht so deutlich wie bei anderen Lexika. Vereinzelt erreichen auch Verfahren ohne diese ansprechende Leistungen und Verfahren mit diesen weniger gute Erkennungsraten. Bezüglich Lemmati-

sierung kann kein besonderer und konsistenter Zusammenhang identifiziert werden. Insgesamt ist keine Lemmatisierung und die Lemmatisierung mittel patter-Lemmatisierer der textblob-Library am effektivsten. Da das GPC, ähnlich zu SentiWS, jedoch ohnehin auch flektierte Wortformen enthält ist der mangelnde deutliche Einfluss von Lemmatisierung, da alle Wortformen ohnehin vorliegen, erwartungskonform. Mit der erweiterten gefilterten Stoppwortliste werden die besten Ergebnisse erzielt.

Das beste Verfahren verwendet eine Lexikonerweiterung mit Lemmatisierung auf Textebene, die Stoppwortliste enhancedFilteredList und Beachtung der Groß- und Kleinschreibung. Dieses Verfahren erkennt 44 von 61 positiven Repliken korrekt und 77 von 139 negative korrekt. Über Analyse von Recall und Precision lässt sich eine übermäßige Zuweisung von positiven Klassen erkennen. Dies führt zu 62 false positives, bei nur 17 false negatives. Der hohe Recall bei positiven Repliken führt aber zu insgesamt guten F-Werten für diese Klasse. Der Unterschied zwischen dem Recall für positive und dem Recall für negative Repliken ist zwar deutlich aber nicht so stark ausgeprägt wie bei anderen Lexika.

Insgesamt weist die Metrik polarityGpc nach Filterung von problematischen Methoden doch eine zufriedenstellende Erkennungsrate über der random baseline auf und kann demnach gemäß der eigenen Evaluationskriterien als brauchbar für die Dramenanalyse angesehen werden.

8.3.3.6 Kombiniertes Lexikon

Zuletzt werden nun noch die Metriken betrachtet, die durch die Kombination von Lexika erstellt wurden: polarityCombined und clearlyPolarityCombined. Die genaue Definition und Erstellung dieser Metriken wurde in Kapitel 5.3.2 besprochen. Über polarityCombined werden alle Wörter aller Lexika zusammengefasst. Ambiguitäten werden durch Mehrheitsentscheidung gelöst oder durch dasjenige Lexikon, das gemäß vorliegender Evaluation eine bessere Erkennungsrate gezeigt hat. Über die Metrik clearlyPolarityCombined werden nur Wörter genutzt die eine eindeutige Mehrheit von 3 Lexika bezüglich einer Polarität aufweisen, wodurch also nur Wörter aufgenommen werden, die in mindestens 3 Lexika enthalten sind und dabei eindeutige Polarität aufweisen. Folgende Durchschnittswerte lassen sich für polarityCombined konstatieren:

Tabelle 43: Übersicht Evaluationsmaße polarityCombined

Evaluationsmaß	Minimum	Durchschnitt	Maximum
accuracy	0,4	0,484	0,675
F-Measure Positive	0,4102564103	0,474	0,524
F-Measure Negative	0,3058823529	0,470	0,773

Die Metrik weist im Schnitt über alle Herangehensweisen eine mittlere Erkennungsrate auf, jedoch mit einem zufriedenstellenden Maximum von 0,675 für die beste accuracy. Diese liegt über der random baseline und lediglich knapp unter der majority baseline (0,695). Die F-Maße lassen erkennen, dass die Leistung für negative Repliken starken Schwankungen, je nach Herangehensweise, unterworfen ist, jedoch im besten Fall mit einem Wert von 0,77 sehr gut ist. Die Prädiktionsleistung für positive Repliken ist hingegen insgesamt schlechter und aufgrund des geringen Maximums erkennt man, dass die Metrik für diese Gruppe Probleme bei der Vorhersage hat. Es werden nun noch die besten sowie eine der schlechteren Herangehensweisen betrachtet:

Tabelle 44: Ausschnitt Evaluationsmaße Verfahrenskombinationen polarityCombined

DTAExtension	Lemmatizer	Lemmatization Type	Stopwords	Case Sensitivity	Accuracy	F-Measure Positive	F-Measure Negative
dtaExtended	treetagger	textLemma	noStopwordList	caseInSensitive	0,675	0,444	0,770
dtaExtended	textblob	textLemma	noStopwordList	caseInSensitive	0,675	0,424	0,773
dtaExtended	textblob	textLemma	noStopwordList	caseSensitive	0,675	0,424	0,773
dtaExtended	tokens	noLemma	noStopwordList	caseInSensitive	0,67	0,467	0,760
dtaExtended	tokens	noLemma	noStopwordList	caseSensitive	0,665	0,446	0,759
…	…	…	…	…	…	…	…
dtaExtended	textblob	bothLemma	standardList	caseSensitive	0,4	0,469	0,310

Wie bereits bei allen bisherigen Metriken erkennt man in der Gesamttabelle deutlich, dass die DTA-Erweiterung die Ergebnisse am meisten verbessert. Bezüglich Lemmatisierung lässt sich kein konsistentes Muster erkennen. Verfahren erbringen auch ohne Lemmatisierung gute Ergebnisse, die Lemma-Erweiterung mittels Lexikon-

Erweiterung ist ausreichend. Analog zu SentiWS und GPC kann man wieder feststellen, dass die 20 besten Verfahren allesamt keine Stoppwortlisten verwenden.

Das beste Verfahren verwendet die Lexikonerweiterung und eine Lemmatisierung auf Textebene mittels Treetagger ohne auf Groß- und Kleinschreibung zu achten. Somit werden 135 Repliken korrekt erkannt. Wie obige F-Werte jedoch zeigen, ist die Erkennung von negativen Repliken deutlich effizienter. Es werden 109 von 139 korrekt erkannt und lediglich 26 von 61 positiven Repliken. Eine deutliche Überklassifikation negativer Repliken wie bei GPC kann jedoch nicht identifiziert werden. Es werden etwa gleich viele false positives (30) wie false negatives (35) produziert.

Aufgrund der Auffälligkeit, dass Verfahren ohne Stoppwortlisten insgesamt deutlich bessere Evaluationsmaße aufweisen, wurde dieser Umstand auf Repliken- und Wortebene explizit inspiziert. Da es sich um eine Metrik eines kombinierten Lexikons handelt, sind alle problematischen als SBW erkannten Stoppwörter von SentiWS und GPC auch hier erhalten. Insgesamt ist der Einfluss aber deutlich geringer als bei GPC, da einzelne Wörter aufgrund der großen Menge an Wörtern, die als SBWs identifiziert werden keinen größeren Einfluss auf die Gesamtklassifikation haben. Auch gleichen sich positive Stoppwörter und negative Stoppwörter des GPC wieder teilweise aus. Es lässt sich aufgrund der Menge an Optionen und den Detailunterschieden auch keine konsistente Systematik erkennen. Manche Verfahren ohne Stoppwortlisten neigen auch zur Überklassifikation mit positiven Auszeichnungen. Insgesamt werden aus den genannten Gründen die Methoden ohne Stoppwortlisten trotz Fehlinterpretation einiger Stoppwörter beibehalten. Dies muss jedoch kritisch in der Resultatinterpretation beachtet werden. Zukünftige Studien können sich mit dem Problem von Stoppwörtern und der damit zugehörigen Überklassifikation genauer auseinandersetzen. Das Problem wird auch in der Diskussion in Kapitel 11 nochmal angesprochen.

Zuletzt werden noch die Evaluationsdaten der Metrik clearlyPolarityCombined betrachtet:

Tabelle 45: Übersicht Evaluationsmaße clearlyPolarityCombined

Evaluationsmaß	Minimum	Durchschnitt	Maximum
accuracy	0,38	0,440	0,51
F-Measure Positive	0,285	0,404	0,479
F-Measure Negative	0,338	0,4675	0,585

Es lässt sich deutlich erkennen, dass diese Metrik eine weniger gute Leistung erbringt als polarityCombined. Das Maximum der Erkennungsrate liegt bei 0,51 und ist damit unter der random baseline von 0,525. Die F-Werte belegen auch, dass die Metrik die schlechtere der beiden Kombinations-Metriken ist. Die bessere Leistung bezüglich negativer Repliken bleibt bestehen. Zur Vollständigkeit werden auch hier noch die besten Verfahren tabellarisch aufgelistet:

Tabelle 46: Ausschnitt Evaluationsmaße Verfahrenskombinationen clearlyPolarityCombined

DTAExtension	Lemmatizer	Lemmatization Type	Stopwords	Case Sensitivity	Accuracy	F-Measure Positive	F-Measure Negative
dtaExtended	textblob	bothLemma	enhancedList	caseSensitive	0,51	0,443	0,562
dtaExtended	treetagger	bothLemma	enhancedList	caseInSensitive	0,505	0,385	0,585
dtaExtended	treetagger	bothLemma	enhancedList	caseSensitive	0,505	0,385	0,585
dtaExtended	treetagger	bothLemma	enhancedFilteredList	caseSensitive	0,505	0,459	0,543
dtaExtended	textblob	bothLemma	enhancedList	caseInSensitive	0,505	0,440	0,556
…	…	…	…	…	…	…	…
noExtension	tokens	noLemma	standardList	caseInSensitive	0,38	0,340	0,415

Grundsätzlich kann man eine Verbesserung der Leistung durch alle Optionen erkennen: Lexikonerweiterung, beidseitige Lemmatisierung und Einsatz von erweiterten Stoppwortlisten. Das Problem der Stoppwörter kann sich hier nicht manifestieren, da diese sich nur in vereinzelten Lexika befinden und clearlyPolarityCombined nur Wörter als SBWs erkennt, die auch in mehreren Lexika stabil enthalten sind.

Das beste Verfahren nutzt Lexikonerweiterung, beidseitige Lemmatisierung, eine erweiterte Stoppwortliste und die Beachtung von Groß und Kleinschreibung. Dadurch

können 102 Repliken korrekt erkannt werden mit einer Überklassifikation positiver Repliken (false positives 76). Im Vergleich ist das Verfahren jedoch bezüglich der meisten Maße schlechter als die vorige Metrik und hat vor allem Problem bei der effektiven Prädiktion negativer Repliken.

Die Metrik polarityCombined weist insgesamt gute Ergebnisse für das hier verwendete Dramenkorpus auf. Die Besonderheit in der Interaktion von Stoppwörtern muss jedoch beachtet werden.

8.3.4 Evaluationsergebnisse – Metriken im Vergleich

Die einzelnen Evaluationsergebnisse der Metriken werden nun in einen vergleichenden Gesamtkontext gestellt um die besten Maße zur Integration in das spätere Front-End zu identifizieren.

Im Anhang findet man hierzu Tabellen aller 800 Verfahren auch geordnet nach accuracy. Zur besseren Übersicht wird auch eine nach accuracy geordnete Tabelle der je 5 besten Verfahren jeder Metrik mit allen Evaluationsdaten mitgereicht, über die man Lexikon-übergreifende Zusammenhänge feststellen kann. Die wichtigsten Erkenntnisse werden an dieser Stelle jedoch nur grob anhand vereinzelter Visualisierungen zusammengefasst.

Folgendes Balkendiagramm illustriert die jeweils besten Verfahren gemäß accuracy pro Metrik. Für polaritySentiWS werden zwei Verfahren angegeben (polaritySentiWS-1 und polaritySentiWS-2), die beide in der Gesamtinterpretation Vor- und Nachteile aufweisen (siehe Kapitel 8.3.3.1). Für die Metrik polarityGpc wurden fehlerbehaftete Verfahren, die oberflächlich betrachtet bessere Ergebnisse produzieren, aus der Auswertung entfernt.

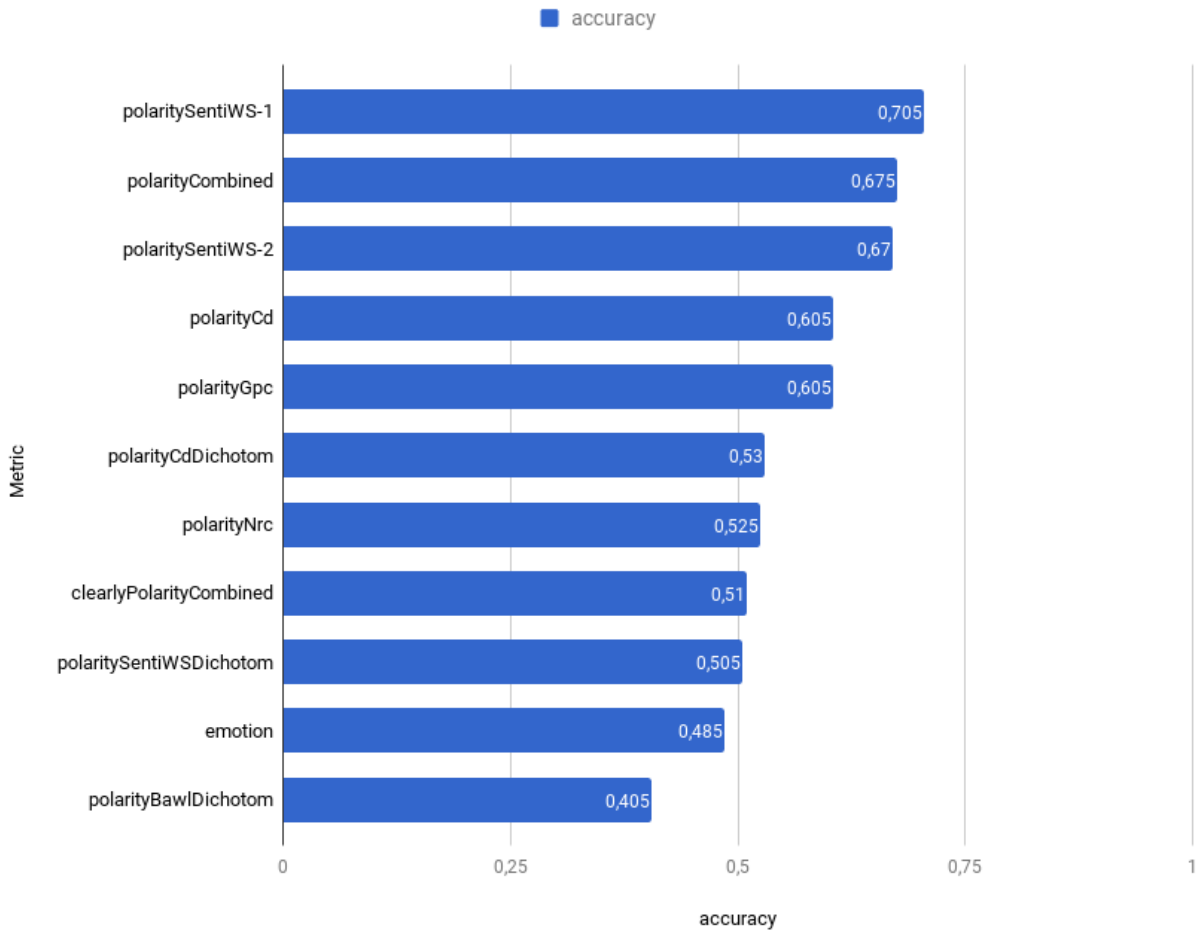


Abbildung 35: Accuracy im Vergleich pro Metrik

Man erkennt deutlich, dass polaritySentiWS-1 die beste Erkennungsrate vorweist. Es werden mit der besten Vorgehensweise von polaritySentiWS 25 von 61 positiven Repliken und 116 von 139 negativen Repliken. Generell kann man bei Analyse der ausführlichen Tabelle erkennen, dass diese Polaritätsmetrik viele der besten Genauigkeiten produziert. Selbst das schlechteste Verfahren bezüglich accuracy ist mit einer Rate von 0,53 besser als einige der besten hier aufgeführten Verfahren. Die Metrik polaritiySentiWS ist auch die einzige die die Benchmark der majority baseline überschreitet (0,695). Auch das sechstbeste Verfahren bezogen auf polaritySentiWS (polaritySentiWS-2) weist noch immer die drittbeste Leistung insgesamt auf. Es wird weiter unten noch diskutiert warum dieses Verfahren insgesamt vielversprechender sein könnte als polaritySentiWS-1. Sehr deutlich ist jedoch, dass die dazugehörige Term-Zähl-Variante polaritySentiWSDichotom sehr schlecht evaluiert wurde und das beste Verfahren noch unter der random baseline liegt. Die Polaritätsstärken von SentiWS konnten damit als

sehr gewinnbringend und notwendig für eine korrekte SA auf dem vorliegenden Korpus identifiziert werden. Insgesamt ist SentiWS gemäß der Evaluationskriterien das am besten geeignete SA-Lexikon.

Zweitbeste Metrik gemäß der accuracy ist mit einem Wert 0,675 polarityCombined. Die Lexikonkombination hat sich damit als durchaus nützliches Verfahren erwiesen, kann jedoch nicht ganz die majority baseline erreichen. Die ebenso auf Basis einer Lexikonkombination entwickelte Metrik clearlyPolarityCombined ist deutlich schlechter und liegt auch mit dem besten Verfahren unter der random baseline.

Grundsätzlich ist die Mehrzahl der Ergebnisse zufriedenstellend, wenn man die random baseline als Benchmark betrachtet. Fast alle Lexika liegen mindestens über der random baseline. Die Metriken von CD und GPC sind etwa gleichwertig. Für CD fällt auch auf, dass sich die Metrik über Polaritätsstärken bezüglich der Genauigkeit gegenüber dem reinen Term-Zähl-Verfahren (polarityCdDichotom) durchsetzt. Der Unterschied ist jedoch nicht so groß wie bei SentiWS, da es nur 2 Ausprägungen von Polaritätsstärken im CD gibt (0,7 und 1). Das NRC liegt mit der besten Methode exakt auf der random baseline. Die Metriken des BAWL-R erzeugen die schlechtesten Erkennungsraten, was an einer übermäßigen Zuteilung von positiven Klassen liegt. Dies führt aufgrund des viel größeren Anteils negativer Repliken im Gesamt-Korpus zu sehr schlechten Prädiktionsergebnissen. Die Metriken des BAWL-R, polaritySentiWSDichotom und clearlyPolarityCombined liegen unter der random baseline. Ihre Prädiktion ist damit schlechter als der Zufall und die Verwendung für die Dramenanalyse kann auf Basis der vorliegenden Evaluation nicht empfohlen werden.

Die auf der Übereinstimmung von Annotatoren basierenden Benchmarks können nur zum Teil erreicht werden. Nutzt man das Maß K-Alpha als Benchmark, erkennt man, dass fast alle Metriken diesen Wert erreichen (0,47). Da K-Alpha unter der random baseline liegt, wird diese Benchmark jedoch hier nicht weiter betrachtet. Die prozentuale Übereinstimmung von 0,765 wird von keinem Verfahren mit keiner Metrik erreicht. Dazu müssten 153 Repliken korrekt erkannt werden, das Maximum erreicht polaritySentiWS mit 141 korrekten Repliken. Die majority baseline wird mit dieser Methode jedoch erreicht.

Tatsächlich wird nun aber zur finalen Umsetzung im Front-End polaritySentiWS-2 gewählt. Die Gründe hierfür wurden bereits in Kapitel 8.3.3.1 angesprochen. Die Met-

rik weist eine ähnlich gute Gesamtleistung wie polaritySentiWS-1 und polarityCombined auf, ist jedoch deutlich besser bei der Erkennung positiver Repliken. Die Methodik weist auch die besten durchschnittlichen F-Werte auf (F-Measure Positive = 0,53; F-Measure Negative = 0,75), was zeigt dass dieses Verfahren mit dieser Metrik eine sehr ausgeglichene Leistung bezogen auf positive und negative Repliken aufweist. Nach Analyse aller Verfahren hat man sich deswegen dazu entschieden, trotz der schwächeren accuracy dieses Verfahren für das Front-End einzubauen. Eine einfache Ersetzung für das Front-End wurde implementiert und wird noch im folgenden Kapitel 9 besprochen.

8.4 Diskussion und Fazit

Analog zu Kapitel 7 werden die Ergebnisse der Evaluationsstudie separat interpretiert und diskutiert. Die wichtigsten Erkenntnisse werden zusammengefasst und in einen Gesamtkontext gestellt. Ferner werden Grenzen des gewählten Verfahrens aufgezeigt und Möglichkeiten zur Verbesserung diskutiert.

Vergleich man die einzelnen Methoden untereinander, dann kann man bezüglich der Auswahl des Lexikons festhalten, dass SentiWS generell und auch bei den meisten sonstigen Kombinationen von Optionen die beste Erkennungsrate aufweist, gefolgt von CD und GPC. Als weiteres Lexikon-spezifisches Resultat fällt stets auf, dass die Verwendung von Polaritätsstärken effektiver ist als reine Term-Zähl-Verfahren, falls Polaritätsstärken vorliegen. Dieser Zusammenhang gilt grundsätzlich meist unabhängig von gewählter Option, wenn für alle Lexika die gleiche Option gewählt wird. Das BAWL-R hat sich aufgrund seiner Übergewichtung positiver Prädiktionen als untauglich gemäß dem gewählten Evaluationsverfahren erwiesen. Das heißt in der Tat, dass je größer das Lexikon im Ausgangszustand ist, desto besser die Erkennungsrate. SentiWS ist zwar bezogen auf die Grundformen eher klein, jedoch aufgelöst das Größte der Lexika (32 734 Einträge). Die Lexikonkombination erzielt ebenfalls gute Ergebnisse für eine Metrik (polarityCombined), ist jedoch in insgesamt für die meisten Evaluationsmaße schlechter als SentiWS generell. Die Ausweitung des Gesamtwortschatzes über diese Methode führt gleichzeitig zu einer Über-Erkennung von SBWs, was im Vergleich zu dem Einzellexikon SentiWS zu einer weniger präzisen Erkennung führt. Auffällig am Vergleich der Lexika ist auch, dass Lexika mit manuell angegebenen Flekti-

onsformen insgesamt meist besser sind als Lexika, die nur aus Grundformen bestehen. Dies gilt häufig unabhängig von gewählter Methodik, also auch Lemmatisierung. Dies kann als erstes Indiz betrachtet, werden, dass die manuelle Angabe von Flektionen präziser ist als die Nutzung eines Lemmatisierers.

Betrachtet man nur die Lemmatisierung als Methodik kann man feststellen, dass diese meist zu einer Verbesserung aller Maße führt, wenn keine DTA-Erweiterung ausgeführt wurde, insbesondere und trivialerweise bei Lexika ohne Flektionsformen wie CD und NRC. Lemmatisierung unterstützt aber auch bei Lexika mit Flektionsformen den Wortabgleich, jedoch in geringerem Ausmaß. Beide Lemmatisierer sind meist ähnlich gut. Es lassen sich nur bei detaillierter Analyse Lexikon-spezifische größere Unterschiede feststellen. Als Lemmatisierungstyp fällt auf, dass textLemma, also die Lemmatisierung des Textes ohne Lemmatisierung des Lexikons meist am besten ist. Die Unterschiede zu lemmaBoth sind jedoch meist gering. Informell sei festzuhalten, dass beide Lemmatisierer fehlerhafte Grundformen produzieren, und was das Hauptproblem für die Methodik lemmaBoth ist, dass die Lemmatisierung im Satz oft anders verläuft als für ein einzelnes Wort.

Als die zentrale Optimierungsmethode kann die DTA-Erweiterung über das Tool von Jurish (2012) identifiziert werden. Sie verbessert die Leistung für alle Lexika unabhängig von alle anderen Optionen deutlich. Auffällig ist dabei, dass die DTA-Erweiterung, selbst wenn sie als einzige Option genutzt wird, eine sehr viel bessere Gesamtleistung produziert als alle anderen Optionenkombinationen ohne DTA-Erweiterung. Die starke Vergrößerung des lexikalischen Wortschatzes führt also tatsächlich zu einer Verbesserung. Durch die DTA-Erweiterung werden neben orthographischen und phonetischen Varianten auch Lemma-Versionen sowie Flektionsformen übergeben. Die Daten legen wieder nahe, dass derartige manuelle Angaben effektiver sind für den Wortabgleich als die Verwendung eines automatischen Lemmatisierers.

Die Rolle und der Einfluss von Stoppwörtern haben sich als sehr speziell herausgestellt. Die angestrebte Verbesserung der Erkennungsleistung durch verschiedene Stoppwortlisten kann nur in vereinzelten Lexika festgestellt werden, z.B. CD. Dann handelt es sich meist um die mit den häufigsten Wörtern des Korpus erweiterten Listen. Tatsächlich musste man aber feststellen, dass die Beachtung von Stoppwörtern häufig zu schlechteren Erkennungsleistungen geführt hat. Die Gründe hierfür sind von

Lexikon zu Lexikon unterschiedlich und werden im Ergebnisteil ausführlicher besprochen. Im Fall des GPC beispielsweise werden viele Stoppwörter als negativ angegeben, was zu einer erhöhten negativen Prädiktion führt, was wiederum aufgrund des ungleichmäßig verteilten Korpus zu einer fälschlicherweise besseren accuracy führt. Es war nicht immer klar wie man mit derartigen Effekten umgehen sollte, meist war der Einfluss nicht derart deutlich wie beim GPC. Bei der final gewählten Methode führen Stoppwörter zu einer generellen positiven Gewichtung, die insgesamt zu gleichmäßigeren Ergebnissen führt, obschon die tatsächliche SBW-Zuweisung falsch ist. Da es sich bei Stoppwörtern nur um eine von vielen Optionen handelt, kann diese Option und ihre Bedeutung nicht erschöpfend analysiert werden, es wurden heuristische Entscheidungen getroffen. Zukünftige Studien sollten sich im Detail mit dem Problem von Stoppwörtern und der Interaktion mit anderen Methoden und der Erkennungsrate auseinandersetzen.

Bezüglich der Beachtung von Groß- und Kleinschreibung konnte festgestellt werden, dass diese bei der Gesamtanalyse kaum einen Einfluss hat und nur bei detaillierter Betrachtung Fehler oder Verbesserungen auftreten.

Insgesamt ist bei der Interpretation der Daten stets die besondere Annotations-Konstitution des Korpus zu beachten, also vor allem das deutliche Übergewicht an negativ annotierten Repliken. Dies hat dazu geführt, dass Verfahren, die generell zu übermäßiger negativer Prädiktion tendieren (SentiWS, GPC) bessere Erkennungsraten aufweisen, als Lexika die übermäßig positive Prädiktionen durchführen (BAWL-R). Es wurde versucht diesen Befund bei der Interpretation der Daten und bei der Identifikation der besten Methode zu beachten. Nachfolgende Studien können dieses Problem durch die Erstellung eines größeren und gleichmäßig verteilten Korpus umgehen. Das Übergewicht von Negativität im Korpus ist jedoch eine Realität und relativierend sei anzumerken, dass die Verfahren die vermehrt negative Repliken vorhersagen möglicherweise auch durchaus korrekt funktionieren. Ein Vergleich mit einem anderen Anwendungsgebiet wäre vonnöten, um sich dieser Frage anzunehmen.

Als weitere Probleme und Grenzen basierend auf der Zusammenstellung des Test-Korpus sei noch erwähnt, dass dieser bezüglich der Annotation mit zahlreichen gemischten Repliken ausgezeichnet wurde. Teilnehmer mussten sich final für eine Polarität entscheiden; in der Tat wurde jedoch eine relevante Zahl an Repliken als gemischt

und neutral wahrgenommen. Auch ist die Übereinstimmung im Vergleich zu anderen Studien recht gering, das heißt die maschinelle Prädiktion ist grundsätzlich anspruchsvoller und Probleme in der Vorhersage von Repliken, die eigentlich als gemischt oder neutral wahrgenommen, sind erwartbar.

Unter diesem Hintergrund sind die finalen End-Ergebnisse zufriedenstellend. Mit einer accuracy von 0,705 ist die Methodenkombination aus dtaExtended, textlemma, treetagger, noStopwordList und caseInSensitiv. Es wurden also etwa 70% korrekt erkannt, dabei mehr negative als positive Repliken. Es konnten also Ansätze gefunden werden, die besser als die random baseline und die majority baseline sind. Die prozentuale Übereinstimmung von Annotatoren konnte jedoch nicht erreicht werden (0,765). Einige Ansätze des GPC erreichen ähnliche Ergebnisse, aber nur weil generell fast alle Repliken als negativ ausgezeichnet werden; aus diesem Grund wurden diese Ergebnisse gefiltert. Als beste Methode wurde jedoch die Kombination dtaExtended, botLemma, textblob, noStopwordList und caseInSensitive von polaritySentiWS gewählt. Diese hat zwar eine geringfügig schlechtere accuracy mit 0,67 jedoch bessere F-Werte bezüglich der Erkennung von positiven und negativen Repliken und somit eine insgesamt ausgeglichenere Prädiktionsleistung. Vergleich man die Erkennungsraten mit anderen in Metastudien gesammelten Ergebnissen erkennt man, dass die Leistung insgesamt etwa leicht unterdurchschnittlich ist (Tsytsarau & Palpanas, 2012; Ravi & Ravi, 2015). Als gut akzeptierte SA auf anderen Anwendungsfeldern erzielte Prädiktionsleistungen im Bereich von 0,8-1. In Anbetracht der bereits erwähnten Schwierigkeiten bei der Prädiktion literarischer Texte und der Tatsache, dass verhältnismäßig primitive Methoden eingesetzt wurden sind die Ergebnisse vielversprechend.

Zuletzt wird noch angesprochen, dass die Emotionskategorien des NRC in keiner Weise auf Korrektheit evaluiert wurden. In der vorliegenden Studie lag der Fokus auf der Polarität. In der Tat konnte man jedoch feststellen, dass emotionale Kategorien insbesondere bedeutend für die SA auf literarischen Texten ist (Mohammad, 2011). Zukünftige Studien können sich explizit mit diesen Kategorien befassen. In der vorliegenden Arbeit wurden die Emotionskategorien jedoch trotzdem in das Front-End eingebaut um erste informelle Analysen zu erlauben.

Über die SA-Evaluation wurden grundsätzlich nur Repliken betrachtet. Es ist zweifelhaft ob dies die bedeutendste Ebene für die literaturwissenschaftliche Interpre-

tation ist oder ob Ebenen wie Szenen oder Sprecherbeziehungen wichtiger sind. Bezüglich dieser Ebene kann keine Aussage bezogen auf die Prädiktionsleistung gemacht werden. Mittels des entwickelten Front-Ends kann man aber auch untersuchen, welche der implementierten SA-Verfahren besser geeignet für globale Interpretationen jenseits der Replik sind. Ebenso kann der Nutzen der Emotionskategorien des NRC exploriert werden.

9 Front-End – Visualisierung

Als finaler Schritt des Projekts wurde eine Web-Anwendung zur Visualisierung und Einsicht der mit dem Back-End produzierten SA-Metriken implementiert. Die Konzeption und Entwicklung wird im folgenden Abschnitt beschrieben. Abschließend wird die Funktionalität anhand Screenshots der Anwendung erläutert. Das Front-End befindet sich im Anhang ist aber auch online erreichbar[6]. Es wird die Verwendung von Google Chrome für das Front-End empfohlen. Des Weiteren findet man Videos mit beispielhaften Anwendungsszenarien im Anhang.

9.1 Idee und Motivation

Die grundsätzliche Idee des Front-Ends ist es, die im Back-End produzierten SA-Daten für alle Ebenen aufzubereiten und zu visualisieren. Das Front-End wurde als Erweiterung eines bestehenden Tools zur Dramenanalyse Katharsis (Burghardt et al., 2016) implementiert. Es orientiert sich bezüglich Design, Aufbau und der grundsätzlichen Software-Architektur am bestehenden Tool. Es fand keine größere Anforderungsanalyse oder mehrstufige Konzeptionsphase statt. Man hat sich bei der Entwicklung der Metriken und Visualisierung an der bestehenden Forschung orientiert und eigene mathematisch Konzepte sowie Visualisierungen konstruiert, deren potentieller Nutzen naheliegend und begründbar ist. Als weiterer Einfluss sei der explizite formulierte Bedarf an unterschiedlichen Funktionen von an dem Projekt interessierten Personen genannt (Betreuer, beteiligter Literaturwissenschaftler).

Ziel ist die Einsicht und Analyse der entwickelten SA über ein interaktives Online-Tool verfügbar zu machen, um erste SA-Studien durchzuführen. Die bisherige Evaluation war auf einen begrenzten Korpus und die Replikenebene beschränkt. Über das

[6] http://lauchblatt.github.io/QuantitativeDramenanalyseDH2015/FrontEnd/sa_selection.html

Tool können aber auch Analysen jenseits des Test-Korpus auf allen Dramen und Ebenen durchgeführt werden. Auf diese Weise kann die Tauglichkeit der SA mit ersten Fallstudien zu literaturwissenschaftlichen Fragen und bekannten Zusammenhängen durchgeführt werden und die SA somit jenseits der Test-Korpus-Ebene evaluiert und untersucht werden. Ferner kann das fertige Tool als guter Ausgangspunkt zur Erhebung und Entwicklung weiterer Anforderungen genutzt werden um in Zukunft konkrete literaturwissenschaftliche Fragestellungen zu formulieren zu deren Bestätigung und Klärung die SA beitragen kann.

9.2 Verwendete Metriken

Wie bereits in Kapitel 8.3.4 besprochen, ist die als optimal eingestufte Methodik eine spezielle Verfahrenskombination über die auf dem Lexikon SentiWS basierende SA-Metrik polaritySentiWS. Es handelt sich um die gewichtete Version der SentiWS-Metrik. Neben dieser Metrik wird auch die Metrik polaritySentiWSDichotom genutzt. Der Grund hierfür ist, dass ein reines Term-Zähl-Verfahren notwendig ist zur Berechnung von Wortverteilungen. Um Verwirrung zu vermeiden hat man sich zum Term-Zähl-Äquivalent von polaritySentiWS entschieden, obschon die Metrik eine schwache Erkennungsleistung bei der systematischen Test-Korpus-Evaluation aufwies. Eine andere Term-Zähl-Methode würde jedoch im großen Widerspruch zu Berechnungen von polaritySentiWS stehen und in der Front-End-Verwendung möglicherweise Probleme bei der Interpretation verursachen. Die in Kapitel 8.3.3.1 diskutierten Probleme von polaritySentiWSDichotom sollten aber bei der Interpretation beachtet werden. Die gewählten Metriken können je nach Bedarf aber durch Anpassung einer js-Datei einfach im Front-End ausgetauscht werden, falls man andere gewichtete oder Term-Zähl-Methoden nutzen will. Es ist jedoch notwendig, dass beide in irgendeiner Form vorliegen. Zur detaillierten Analyse werden neben der finalen jeweiligen Polaritäts-Metrik auch die jeweils positive und negative Kalkulation betrachtet, also im Fall von polaritySentiWS auch polaritySentiWSPositive und polaritySentiWSNegative. Die Subtraktion der negativen Metrik von der positiven ergibt dabei stets die finale Polarität. Durch die Werte können die Polaritätsklassen getrennt analysiert werden und Verteilungen dieser präsentiert werden.

Neben den genannten Polaritäts-Metriken werden noch die Emotionskategorien, berechnet über das NRC, in das Front-End integriert. Es konnte in der Literaturrecherche konstatiert werden, dass Emotionskategorien für zahlreiche literaturwissenschaftliche Fragestellungen von Bedeutung sind. Auch in der Test-Korpus-Annotation können vereinzelte Emotionen als häufig auftretenden Assoziationen mit Repliken identifiziert werden. Neben den acht Haupt-Emotionskategorien wird auch die Variable Emotion vorhanden (emotionPresent) in die Analyse aufgenommen. Diese gibt für ein Wort an, ob irgendeine der Emotionen gemäß NRC damit assoziiert wird. Durch die Erweiterung der Polaritäts-Sentiment mit Emotionskategorien kann die Nutzung dieser in der Literaturwissenschaft genauer untersucht werden. Es ist jedoch zu beachten, dass die Emotionskategorien des NRC, im Gegensatz zu den Polaritäts-Metriken, in ihrer Leistung nicht systematisch evaluiert wurden

Im Back-End können die genannten Metriken gemäß dem erwünschten Verfahren berechnet und ausgegeben werden. Dies geschieht auf verschiedenen Ebenen und auf Basis unterschiedlicher Normalisierungsstufen. Diese werden für die Ausgabe im Front-End aufgegriffen. Es wird zwischen der strukturellen Ebene, der Sprecher-Ebene und der Sprecherbeziehungs-Ebene (Charakter-zu-Charakter-Ebene) unterschieden. Die Ebenen wurden bereits in Kapitel 5.5 eingeführt. Sie bilden im Front-End auch die obersten Auswahlpunkte im Menü. Auf der strukturellen Ebene kann man SA-Metriken für das Drama als Ganzes, den Akt, die Szene und auf unterster Ebene der Replik einsehen. Auf der Sprecher-Ebene kann man Sprecher-spezifische SA-Metriken pro Drama, Akt, Szene und Replik einsehen und Sprecher-Metriken untereinander vergleichen. Sprecherbeziehungen können auf Dramenebene, pro Akt und pro Szene betrachtet werden. Ferner können Metriken auf drei Arten normalisiert sein. Die Metriken können absolut, normalisiert an der Anzahl aller Sentiment-Tragender Wörter oder normalisiert an der Anzahl der Wörter betrachtet werden. Die Wortanzahl bezieht sich dabei immer auf das Bezugsobjekt (also z.B. den Akt, die Sprecherrepliken per Szene usw.). Die Normalisierung der Sentiment-Wörter bezieht sich immer nur auf die einer expliziten Metrik zugeordneten Zahl an Wörtern, also für polaritySentiWS nur die für die Kalkulation zugehörigen SBWs und für die NRC-Emotionen auch nur die SBWs, die zur Kalkulation der NRC-Emotions-Metriken beitragen

Als weitere wichtige Dimension zum Einsatz der Metriken wird im Front-End meist noch zwischen dem dynamischen Verlauf und der statischen Verteilung unterschieden. Der dynamische Verlauf erlaubt die Betrachtung von Metriken über strukturelle Einheiten hinweg, also die Ausprägung von Metriken über Akte, Szenen und Repliken hinweg, während bei der statischen Verteilung Häufigkeitsverteilungen von SBWs auf Textebene betrachtet werden. Für letztgenanntes Verfahren werden Metriken nicht isoliert betrachtet sondern die Polaritäts- und Emotionsklassen in ihrer Verteilung, also z.B. die Zahl positiver Wörter im Vergleich zu negativen.

Es bleibt zu beachten, dass das SA-Verfahren des Back-Ends, das im Front-End genutzt wird angepasst werden kann. Hier werden die Namen der standardmäßig gewählten SA-Option angegeben. Diese können in einer Datei des Front-Ends einfach manipuliert und angepasst werden (siehe Kapitel 9.3). Die Namen des Front-Ends sind generisch gewählt und werden deswegen nicht angepasst bei Änderungen

Die Metriken aus dem Back-End wurden in für das Front-End geeignete verständliche Namen übertragen. Folgende Tabellen fassen die einzelnen Metriken mit ihren neuen und alten Namen für die einzelnen Betrachtungsebenen zusammen. Zunächst die Metriken und Normalisierungsoptionen die bei Grafiken auf der Verlaufsebene betrachtet werden können.:

Tabelle 47: Back-End-Metriken übertragen auf Front-End-Namen

Metrik – Back-End-Name	Metrik – Front-End-Name
polaritySentiWS	Polarität (gewichtet)
positiveSentiWS	Positiv (gewichtet)
negativeSentiWS	Negativ (gewichtet)
polaritySentiWSDichotom	Polarität (Wortanzahl)
positiveSentiWSDichotom	Positiv (Wortanzahl)
negativeSentiWSDichotom	Negativ (Wortanzahl)
anger	Zorn
anticipation	Erwartung
disgust	Ekel
fear	Angst
joy	Freude
sadness	Traurigkeit
surprise	Überraschung
trust	Vertrauen
emotionPresent	Emotion vorhanden

Die Metriken sind im JSON unter dem Eintrag sentimentMetricsBasic in verschiedenen Normalisierungslisten gespeichert (siehe nächste Tabelle) und werden mit den hier angegebenen Begriffen bereitgestellt. Einige der Metriken basieren auf Kalkulationen über Term-Zähl-Verfahren (z.B. Polarität (Wortanzahl), Zorn), andere auf Basis von Polaritätsstärken, nämlich alle als gewichtet benannte Metriken. Folgende Normalisierungsoptionen gibt es für die Verlaufsebene.

Tabelle 48: Back-End-Normalisierungen übertragen auf Front-End-Namen

Normalisierung – Back-End-Liste/Name	Normalisierung – Front-End-Name
metricsTotal	Absolut
metricsNormalisedSBWs	Normalisiert an Sentiment-Tragenden Wörtern
metricsNormalisedLenghtInWords	Normalisiert an Anzahl aller Wörter

Im Front-End, dass noch in Kapitel 9.4 ausführlicher beschrieben wird, kann man auf der Verlaufsebene zwischen drei Normalisierungsstufen wählen, die äquivalent zu Listen aus dem Back-End sind, in denen die jeweiligen Metriken normalisiert oder nicht normalisiert gespeichert sind. Diese Listen werden auch im JSON unter dem Eintrag sentimentMetricsBasic gesichert. Es können absolute Werte, an der Zahl an zur Metrik gehörenden SBWs normalisiert Werte, und an der Länge der Anzahl aller Wörter normalisierte Werte betrachtet werden. Handelt es sich bei einer Metrik um ein durch Term-Zähl-Verfahren erstelltes, dann gibt die SBW-Normalisierung den Anteil einer Klasse an zu dieser Metrik gehörenden SBWs wider. Selbiges gilt für die Normalisierung an der Anzahl der Wörter, hier wird der Anteil einer Klasse, gemessen über die Wörter am Gesamttext angegeben. Diese mathematische Annahme ist nicht für finale Polaritäts-Metriken gültig, da diese durch Subtraktion ihrer Klassen entstehen. Aber für folgende Metriken: Positiv (Wortanzahl), Negativ (Wortanzahl), die Emotionskategorien, Emotion vorhanden. Auf diese Weise werden Werte auch bezogen auf die Untersuchungseinheit vergleichbar gemacht. So können durch Normalisierungen Werte vergleichbar gemacht werden, da der Einfluss einer stärkeren Metrik-Ausprägung durch die Textlänge relativiert wird. Ist beispielsweise ein Akt länger bezüglich Wortanzahl, so kann dieser potentiell höhere Werte erhalten. Um Akte nun besser ohne Einfluss dieser Wortanzahl zu vergleichen, werden die Werte über die Normalisierung durch die SBW-Anzahl oder Wortanzahl geteilt. Die Werte repräsentieren also eine Form von Sentiment-Dichte und ermöglichen die Vergleichbarkeit unterschiedlich langer Texte. Normalisierungen können demnach auch für andere Metriken sinnvoll sein, obschon nicht der Anteil dadurch angegeben wird. Für folgende Maße gilt, dass sie zwar keine Anteile durch die Normalisierung angeben, aber die Vergleichbarkeit ermöglicht wird: Polarität (gewichtet), Positiv (gewichtet), Negativ

(gewichtet) Polarität (Wortanzahl). Bei der Normalisierung von Emotion vorhanden ist zu beachten, dass die Normalisierung an den SBW stets 100% ergibt, da alle Emotions-SBWs des NRC trivialerweise eine vorhanden Emotion aufweisen. Es wird also die Zahl an Wörtern bei denen eine Emotion vorhanden ist durch die Zahl aller Emotionswörter geteilt, was stets 1 ergeben muss.

Neben der Verlaufsebene gibt es auch noch eine Verteilungsebene, die im Front-End in unterschiedlichen Visualisierungen genutzt wird. Dabei werden verschiedene Klassen einer Sentiment-Gruppe anteilsmäßig angegeben. Die Sentiment-Gruppen sind dabei die folgenden:

Tabelle 49: Sentiment-Gruppen im Front-End

Sentiment-Gruppe/Name im Front-End	Zugehörige Sentiment-Metriken aus dem Back-End
Polarität (gewichtet)	positiveSentiWS, negativeSentiWS
Polarität (Wort)	positiveSentiWSDichotom, negativeSentiWSDichotom
Emotionen	anger, anticipation, disgust, fear, joy, sadness, surprise, trust
Emotion vorhanden	emotionPresent

Die Kalkulation der Anteile wird im Front-End durch Übergaben der absoluten Werte der einzelnen Metriken einer Gruppe durchgeführt. Diese werden dann in einem Kreisdiagramm illustriert. Es gibt hierbei zwei Verteilungsarten, die man auswählen kann. Die Verteilung von Sentiment-Tragenden Wörtern und die Verteilung von allen Wörtern. Bei erstgenannten werden nur die Anteile der Sentiment-Klassen an allen Sentiments einer Gruppe dargestellt. Bei der Verteilung von allen Wörtern werden die Sentiment-Klassen in ihrem Anteil an allen Wörtern illustriert, also noch eine Klasse von Sentiment-freien Wörtern hinzugezählt. Für Polarität (gewichtet) liegt, da es sich um Polaritätsstärken und kein Term-Zähl-Verfahren handelt, ein Sonderfall vor. Es werden also keine Wortzahl-basierten Verteilungen angezeigt sondern Verteilungen bezüglich Polaritätsstärken. Für Polarität (gewichtet) wird ferner die Verteilung von beiden Verteilungsarten als gleich angezeigt, da es mathematisch unklar ist, wie mit Sentiment-freien Wörtern bei der Verteilungskalkulation umzugehen ist.

9.3 Entwicklung

Das Front-End wurde als SA-Erweiterung des bestehenden Tools Katharsis (Burghardt et al., 2016) als Web-Anwendung mittels HTML, Javascript und CSS implementiert. Zur optimierten Javascript-Entwicklung wurde jQuery verwendet. Zur Darstellung der Grafiken und für einige Interaktionseffekte wurde die Bibliothek Google Chart[7] integriert. Diese bietet zahlreiche Optionen für Visualisierungen von Daten und wurde deswegen für den vorliegenden Anwendungsfall als besonders geeignet angesehen. Des Weiteren wurde dieselbe Library bereits in der ersten Version von Katharsis erfolgreich verwendet und aus Konsistenzgründen ist die Weiterverwendung naheliegend.

Es gibt drei Hauptkomponenten sowie einen Start/Auswahl-Screen. Für alle vier Bestandteile liegen dementsprechend vier HTML-Seiten vor, die für die SA-Komponente deklariert wurden: sa_selection.html für die Auswahl, sa_dramaActsScenes für die strukturelle Analyse, sa_speakers.html für die Sprecher-Analyse und sa_relations für die Sprecher-Beziehungen-Analyse. Zu jeder dieser Komponente liegt ein Paket an js-Dateien vor (Ordner: SA_Selection, SA_ActsScenes, SA_Speakers, SA_Relations). Die Integration der SA-Daten aus dem Back-End wird über eine JSON-Datei, die im Back-End produziert wird, ermöglicht. Es wurden JSON-Ausgabe-Dateien mit allen notwendigen Daten für die wichtigsten und besten SA-Algorithmen produziert. Diese sind im Ordner json einsehbar. Standardmäßig ist die oben beschrieben Ausgabe als Datensatz ausgewählt. Man kann jedoch durch einfache Anpassung einiger globaler Dateien (siehe unten) andere auswählen und so weitere Methoden und Lexika für die komplexere Interpretation im Front-End untersuchen. Die json-Datei wird in der html-Datei deklariert und wie ein normales json-Objekt interpretiert.

Für die Selektionsseite liegt nur eine js-Datei vor, in dem die Darstellung der Auflistung implementiert ist. Für die anderen Komponenten wurden js-Dateien nach dem MVC-Modell implementiert. Das Modell implementiert dabei Algorithmen zum Bezug und der Verarbeitung aller notwendigen SA-Daten aus der JSON-Datei. Ziel ist dabei die Umwandlung in eine Datenstruktur, die von der Google Chart-Library verarbeitbar ist. Es wird dabei für jede Visualisierung eine eigene Datenstruktur produziert.

[7] https://developers.google.com/chart/

Dabei wird zum Teil auf für alle Visualisierungen notwendigen Algorithmen in der globalen js-Datei globals.js zurückgegriffen. Die Daten werden dann über einen Controller an die View gesendet. Für logische Einheiten von Visualisierungen pro Seite wird eine View-Datei implementiert. Hier wird die UI-Logik pro Seite umgesetzt sowie das dynamische Rendern der einzelnen Visualisierungen. In der Controller-Datei wird die Interaktion und generelle Funktionalität zwischen den Model- und View-Programmen geregelt, indem diese als Javascript-Objekte genutzt werden. Ferner wurde ein spezielles Namespace-Paradigma für die Javascript-Entwicklung eingesetzt um die Sichtbarkeit einzelner Variablen zu kontrollieren. Controller-Programme enden mit -Controller im Namen, View-Dateien mit -View und Model-Dateien mit -Model.

Einige wichtige Funktionen auf die alle Teilkomponenten zugreifen müssen, wurden des Weiteren über globale Dateien implementiert. Diese befinden sich im Ordner Globals. Über die Datei setChosenDrama.js wird die Sicherung des ausgewählten Dramas umgesetzt. Die Datei global.js nun beinhaltet Methoden zur Transformation normaler Sentiment-Daten einer Sentiment-Einheit der mitgelieferten JSON-Datei in Sentiment-Anteile zur Nutzung bei Visualisierung von Sentiment-Anteilen sowie in eine Datenstruktur für die Visualisierung auf der obersten Dramen-Ebene. Letztgenanntes verläuft anders als herkömmliche Verlaufsgrafiken, weswegen diese Umwandlung notwendig ist. Ferner sind Übersetzungs-Methoden enthalten, die die englischsprachigen Metrik-Namen in die deutschen Versionen für das Front-End umwandeln. Diese Datei muss man bei der Anpassung und Veränderung der genutzten SA-Methodik aus dem Back-End bearbeiten. So muss man lediglich die Namen der Metriken austauschen, die verwendet wurden, z.B. bei einem Wechsel von polaritySentiWSDichotom; diesen Namen mit dem Namen der neuen Metrik austauschen. Alle anderen Dateien passen sich an, die Namen, die im Front-End präsentiert werden, sind generisch gewählt, so dass eine Anpassung nicht notwendig sein muss.

9.4 Funktionalität

Im Folgenden Abschnitt wird nun die Funktionalität des Front-End-Tools abschnittsweise anhand von Screenshots beschrieben. Einige spezielle Besonderheiten bei der Programmierung, die in Kapitel 9.3 noch nicht genauer angesprochen wurden, werden hier noch vereinzelt nachgetragen. Das Tool wird auch nach Abgabe der vorliegenden

Arbeit weiterentwickelt und optimiert. Es ist möglich, dass sich deswegen Grafiken der Online-Version des Tools von den hier gezeigten und auch von den im Anhang vorliegenden Tool unterscheiden.

9.4.1 Allgemeines Design

Das Design der SA-Komponente orientiert sich am Gesamtkonzept von Katharsis und wurde nicht auf besondere Art und Weise verändert. Als CSS-Framework wird wie in Katharsis das CSS-Framework Bootstrap[8] verwendet. Das Framework bietet CSS-Klassen und Funktionen zur vereinfachten, konsistenten und ansprechenden Gestaltung des User Interface. Die sonstigen Design-Elemente für die SA-Komponente werden über die CSS-Datei sentiment-analysis.css implementiert.

Jeder Menüpunkt gliedert sich in über Rahmen klar abgetrennte funktionale Einheiten. Die Bedeutung jeder dieser Einheiten wird über eine Überschrift angegeben. In den meisten Fällen kann man einzelne Grafiken über Interaktion mit Drop-Down-Menüs anpassen und explorieren. Andere besondere Funktionen werden bei der jeweiligen Visualisierung in den nachfolgenden Beschreibungen genauer besprochen.

Bei der Interaktion mit den Menüpunkten werden CSS3-Transitions verwendet. Beim Wechsel zwischen den einzelnen Menüpunkten werden Fade-Effekte mit jQuery umgesetzt. Die Präsentation der einzelnen Seiten benötigt keine Lade-Screens, da die Kalkulation, Datenverarbeitung und Visualisierung in einer angemessenen Geschwindigkeit stattfindet.

9.4.2 Header

Im Header aller Webseiten befindet sich das bereits für Katharsis genutzte Logo, rechts daneben, wenn eine Auswahl getroffen wurde, der Name, das Jahr und der Autor des Dramas. Als Menüpunkte kann man zwischen der Dramenauswahl, der strukturellen Analyse (Drama-Akt-Szenen-Repliken-Analyse), der Sprecher-Analyse und den Sprecher-Beziehungen wechseln. Die momentane Auswahl wird hervorgehoben.

[8] http://getbootstrap.com/

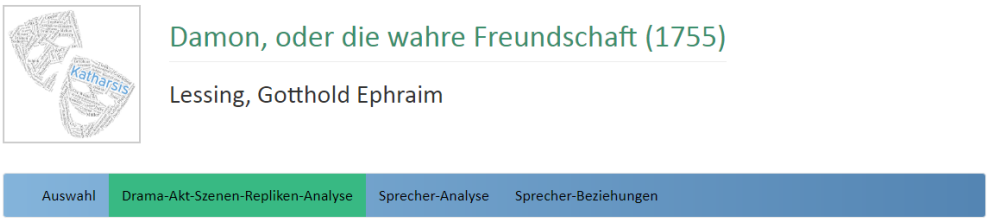


Abbildung 36: Header

9.4.3 Dramen-Auswahl

Folgender Screenshot zeigt die Startseite der SA-Komponente, die Dramenauswahl:

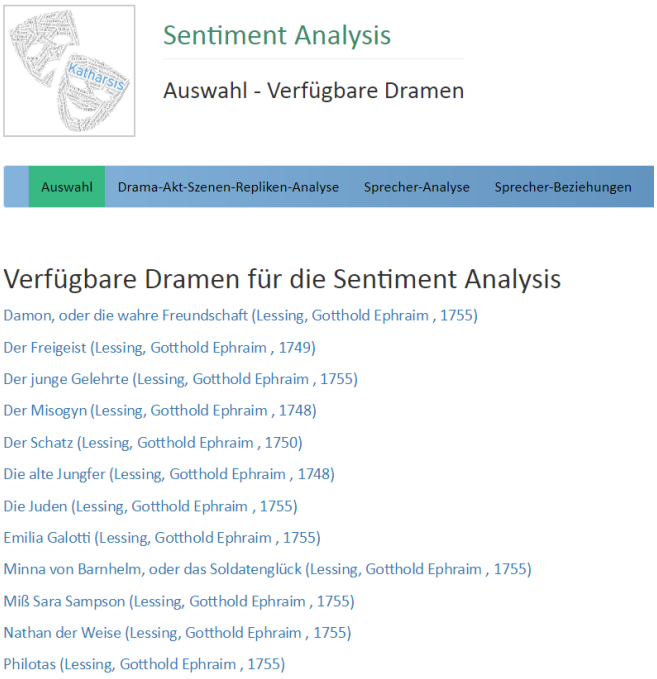


Abbildung 37: Dramen-Auswahl

Die SA-Komponente der entwickelten Katharsis-Erweiterung beginnt mit der Auswahl, der für die SA zur Verfügung stehenden Dramen. Die diesbezügliche html-Datei ist sa_selection.html. Die dazugehörige js-Datei sa_selection.js im Ordner SA_Selection liest dabei lediglich aus der vom SA-Back-End produzierten json-Datei aus, welche Dramen vorliegen, sammelt und strukturiert die Metadaten und visualisiert die Auflistung. Ferner wird das Klick-Event pro Listeneintrag initialisiert, das das momentan ausgewählte Drama sichert. Die momentane Auswahl wird über den localStorage gespeichert und verwaltet.

9.4.4 Strukturelle Analyse

Die Strukturelle Analyse wird im Front-End Drama-Akt-Szenen-Repliken-Analyse genannt um die Bedeutung des Menüpunktes zu vereinfachen und zu verdeutlichen.

Hier können alle Analysen auf struktureller Ebene durchgeführt werden. Als Beispiel-Drama für diese Ebene werden alle Grafiken bezüglich Miss Sara Sampson betrachtet.

9.4.4.1 Sentiments im ganzen Drama

Über die erste Grafik kann man sich Polaritäts- und Emotionswerte auf der obersten strukturellen Ebene also dem Drama in Form eines Balkendiagramms ansehen. Diese werden hier, im Gegensatz zu allen anderen Verlaufsebenen gruppenbasiert angezeigt also pro Polarität (gewichtet), Polarität (Wortanzahl) und Emotionen. Folgender Screenshot zeigt als Beispiel die absoluten Werte von Polarität (gewichtet) für das Drama Miss Sara Sampson:

Abbildung 38: Visualisierung – Sentiments im ganzen Drama (Polarität gewichtet)

Bei der Auswahl von Polaritäten wird die positive Klasse als positive Zahl und grün dargestellt, die negativ als negative Zahl und rot. Die finale Gesamt-Polarität befindet sich in der Mitte und gibt die Gesamtausrichtung des Dramas bezüglich des Polaritäts-Sentiments wider. Hovert man mit der Maus über einen Balken erscheint ein Tooltip mit den exakten Daten für jeden Balken. Über ein Drop-Down-Menü kann man die zu untersuchende Sentiment-Gruppe und die Normalisierung auswählen. Folgender Screenshot zeigt noch beispielhaft die Werte für die Emotionsgruppe, diesmal normalisiert an der Wortanzahl:

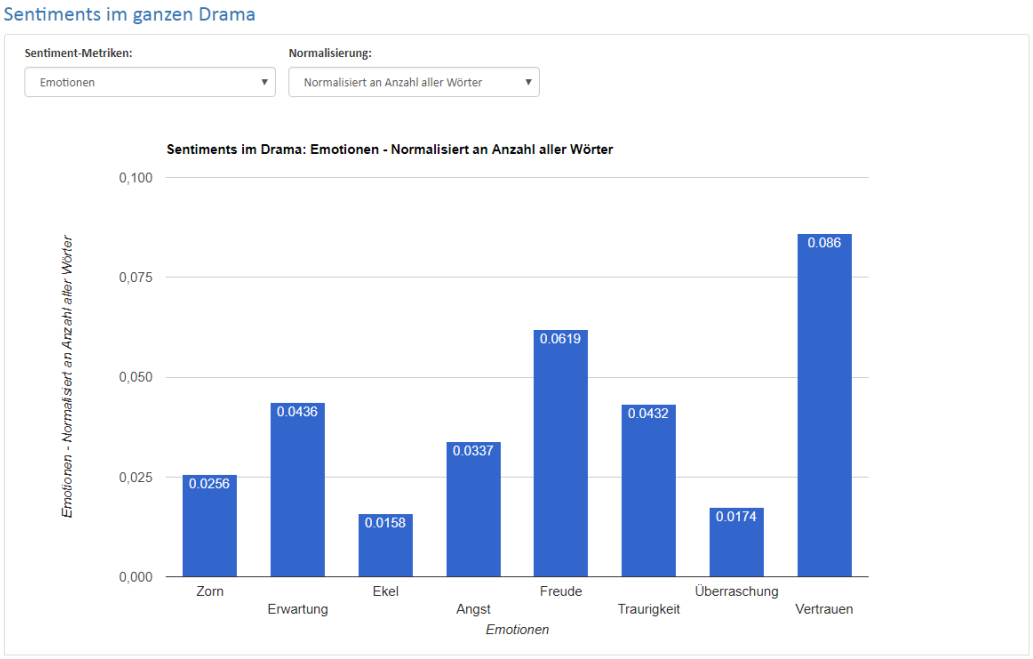


Abbildung 39: Visualisierung – Sentiments im ganzen Drama (Emotionen)

9.4.4.2 Kreisdiagramm – Sentiment-Anteile im Drama

Die zweite Grafik unter diesem Menüpunkt ermöglicht die Visualisierung von Sentiment-Anteilen im Drama. Als Sentiment-Anteile können, wie für jedes derartige Kreisdiagramm im System, Polarität (gewichtet), Polarität (Wortanzahl), Emotionen und Emotion vorhanden, gewählt werden. Polarität (gewichtet) gibt die Sentiment-Verteilung der positiven und negativen Polaritätsstärken im Vergleich wieder, Polarität (Wortanzahl) die Zahl an positiven und negativen Wörtern im Vergleich und Emotionen, die einzelnen Emotionsanteile. Über Emotion vorhanden kann man den Anteil an generell als emotional erkannten Wörtern einsehen. Alle Metriken werden zunächst als Anteil an der Zahl der zur Metrik gehörenden SBWs betrachtet (Drop-Down-Auswahl: Verteilung von Sentiment-Tragenden Wörtern). Es handelt sich also um eine Visualisierung gemäß der Normalisierung an der Zahl an SBWs. Über ein Drop-Down-Menü kann man jedoch auch eine Normalisierung an allen Wörtern auswählen (Verteilung von allen Wörtern), so dass neben den üblichen Sentiment-Kategorien auch eine Kategorie für kein Sentiment-tragende Wörter im Kreisdiagramm aufgeführt wird. Diese Funktion kann nicht auf Polarität (gewichtet) angewendet werden da diese Metrik über Polaritätsstärken verläuft und keine Anteile im Zusammenhang mit nicht Sentiment-tragenden Wörtern berechnet werden können. Es wird stattdessen die Verteilung von Sentiment-Tragenden-Wörtern angegeben.

Folgende Beispieldiagramme visualisieren einmal den Anteil positiver und negativer Wörter an allen Polaritäts-Wörtern und darunter den Anteil an allen Wörtern, um den Unterschied deutlich zu machen:

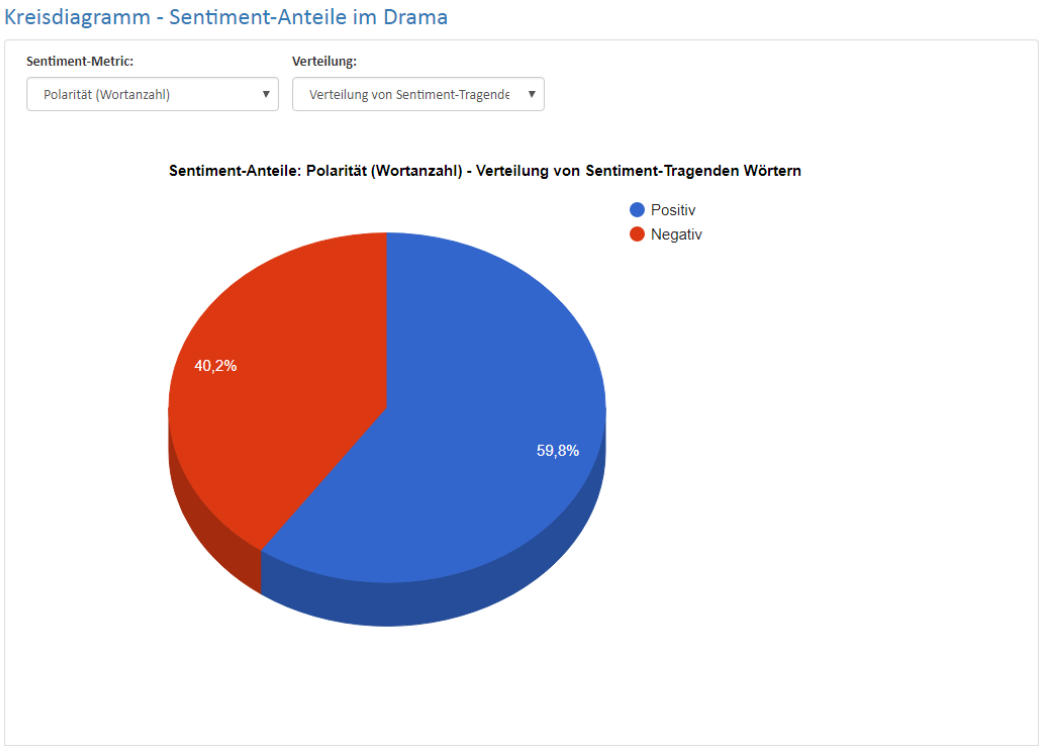


Abbildung 40: Visualisierung – Sentiment-Anteile im Drama I

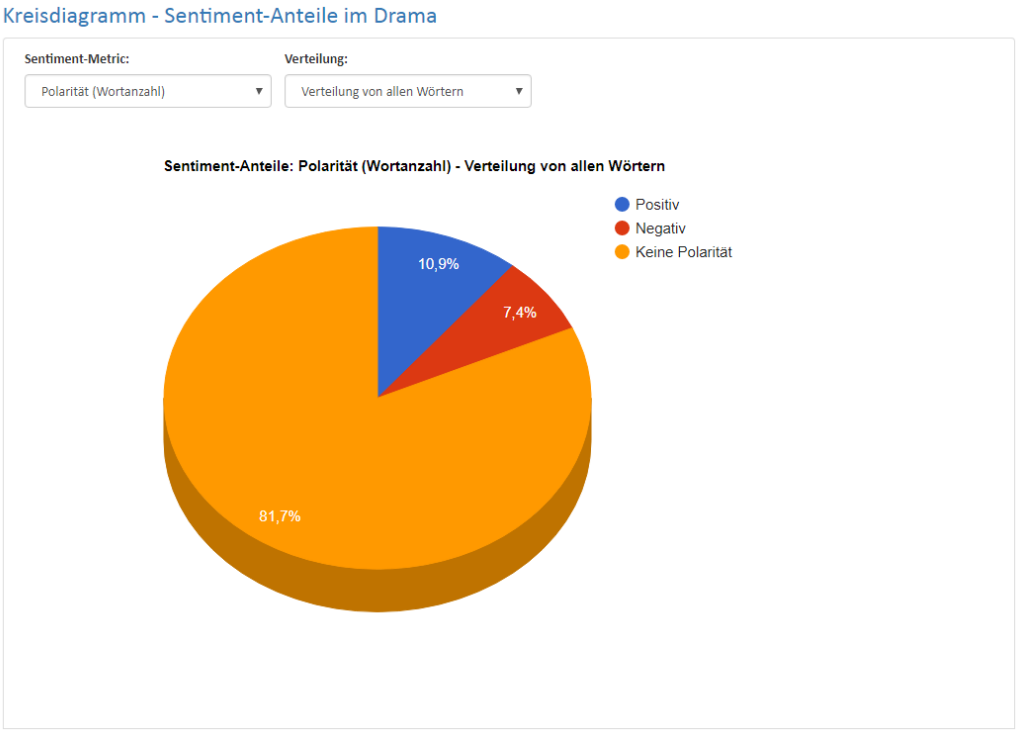


Abbildung 41: Visualisierung – Sentiment-Anteile im Drama II

Hovert man mit der Maus über die einzelnen Anteile, werden die exakten Anteilswerte angezeigt, also im vorliegenden Fall die exakte Zahl an Wörtern.

9.4.4.3 Verlaufsdiagramm – Sentiments im Drama pro Akt

Als erstes Verlaufsdiagramm werden die Sentiments pro Akt über ein Balkendiagramm angegeben. Über ein Drop-Down-Menü kann man wieder zwischen Polarität (gewichtet), Positiv (gewichtet), Negativ (gewichtet), Polarität (Wortanzahl), Positiv (gewichtet), Negativ (gewichtet), den einzelnen Emotionskategorien und Emotion vorhanden wählen. Über ein zweites Drop-Down-Menü kann man die Art der Normalisierung einstellen. Auf der x-Achse wird die Akt-Nummer abgetragen und auf der y-Achse die jeweiligen Sentiment-Werte. Über bzw. in den Balken werden gerundet die jeweiligen Werte pro Akt angezeigt. Auch hier kann man mit der Maus über die Balken gehen, um detailliertere Werte zu erhalten. Folgendes Diagramm zeigt den Polaritätsverlauf für die Metrik Polarität (gewichtet) im Verlauf der Akte von Miss Sara Sampson:

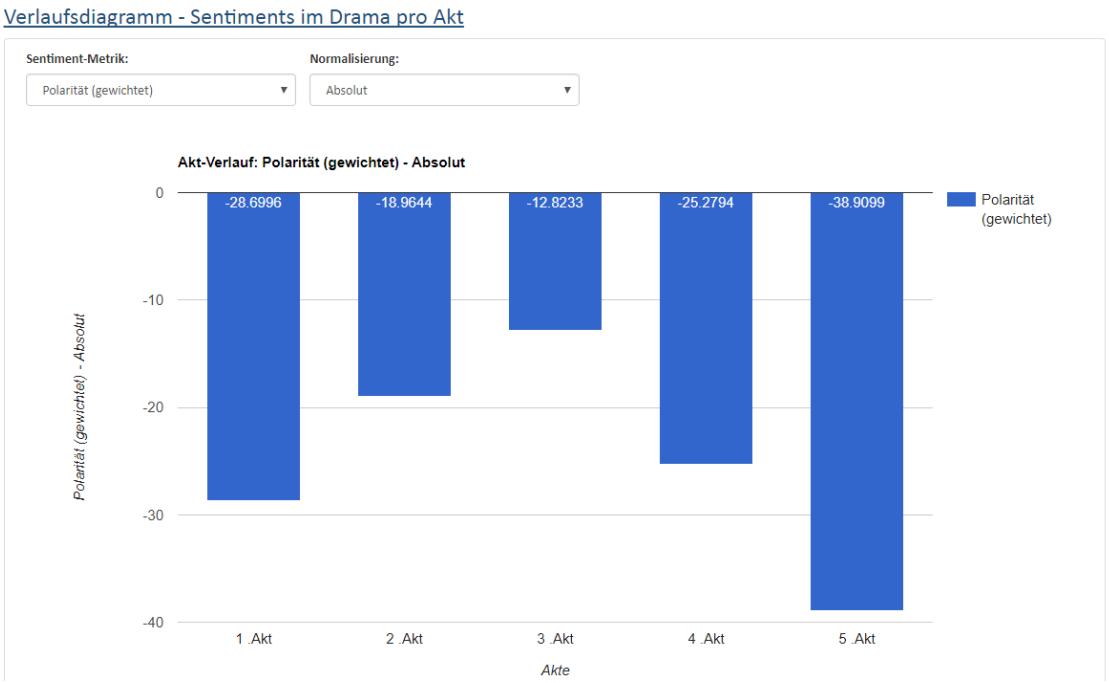


Abbildung 42: Visualisierung – Verlaufsdiagramm – Sentiments im Drama pro Akt

Man erkennt, dass alle Akte negative Gesamt-Polaritäten enthalten, jedoch mit einem Abfall bis zum dritten Akt und einem anschließenden starken Anstieg. Der 5.Akt weist das negativste Sentiment auf. Es ist dabei zu beachten, dass hier die absoluten Werte betrachtet werden, das heißt die Länge eines Aktes hat Einfluss auf den Gesamtwert (da längere Akte potentiell mehr SBWs enthalten können). Um Akte vergleichend zu

betrachten, wird empfohlen, die Normalisierung an der Anzahl an SBWs oder Wörtern zu betrachten.

9.4.4.4 Kreisdiagramm – Sentiment-Anteile pro Akt

Analog zu der besprochenen Visualisierung der Sentiment-Anteile pro Akt in Kapitel 9.4.4.2 kann man die Sentiment-Anteile auch pro Akt einsehen. Die Auswahl der Sentiment-Metrik und der Anteilsart (SBWs oder alle Wörter) verläuft dabei analog. Zusätzlich können jetzt jedoch über ein Drop-Down der jeweilige Akt ausgewählt werden. Folgender Screenshot gibt als Beispiel die Verteilung der Emotionen an allen Emotionskategorien in Akt 5 von Miss Sara Sampson wider:

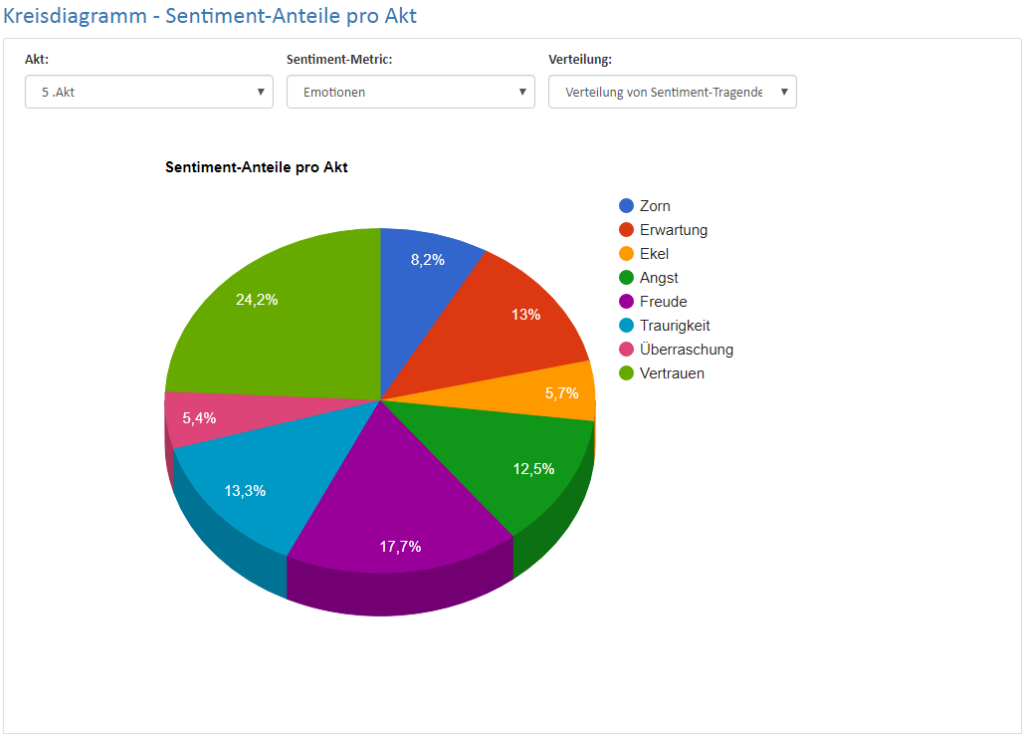


Abbildung 43: Visualisierung – Kreisdiagramm – Sentiment-Anteile pro Akt

Die Sentiment-Anteile von NRC-Emotionswörtern zeigen, dass Vertrauen die häufigste Emotion gemäß dieses Lexikons ist. Indem man über die Kategorie hovert wird über einen Tooltip angezeigt, dass es sich um exakt 397 Wörter handelt. Derartige Visualisierungen mit den NRC-Emotionen wurden bereits bei Mohammad (2011) verwendet, können über das vorliegenden Tool jedoch auf mehr Dramenebenen und mit unterschiedlichen Sentiments betrachtet werden.

9.4.4.5 Verlaufsdiagramm – Sentiments in Szenen pro Akt

Das Verlaufsdiagramm für die Sentiments in Szenen pro Akt ist von der Funktionalität ähnlich aufgebaut wie das Verlaufsdiagramm für die Sentiments pro Akt. Es bestehen

die gleichen Auswahl-Möglichkeiten von Metriken und Normalisierungen wie bei den Sentiments pro Akt. Es wird je eine Grafik für jeden Akt produziert. Auf der x-Achse jeder Grafik ist die Szene abgebildet, auf der y-Achse die jeweilige Sentiment-Metrik. Auf jedem Balken wird der Sentiment-Wert auf 4 Stellen gerundet angegeben. Bei fünf Akten führt das zu fünf Grafiken die untereinander angeordnet sind. Ferner wird eine Trendlinie eingetragen, um den grundsätzlichen Verlauf des Sentiments zu visualisieren. Eine derartige Trendlinie ist zur Interpretation nur bei einer ausreichenden Zahl an Szenen geeignet.

Folgendes Diagramm gibt als Beispiel den Verlauf von Akt 1 bezüglich der negativen Polaritätsgewichte pro Szene wieder. Ferner wurden die Werte an der Anzahl der Wörter in der jeweiligen Szene normalisiert:

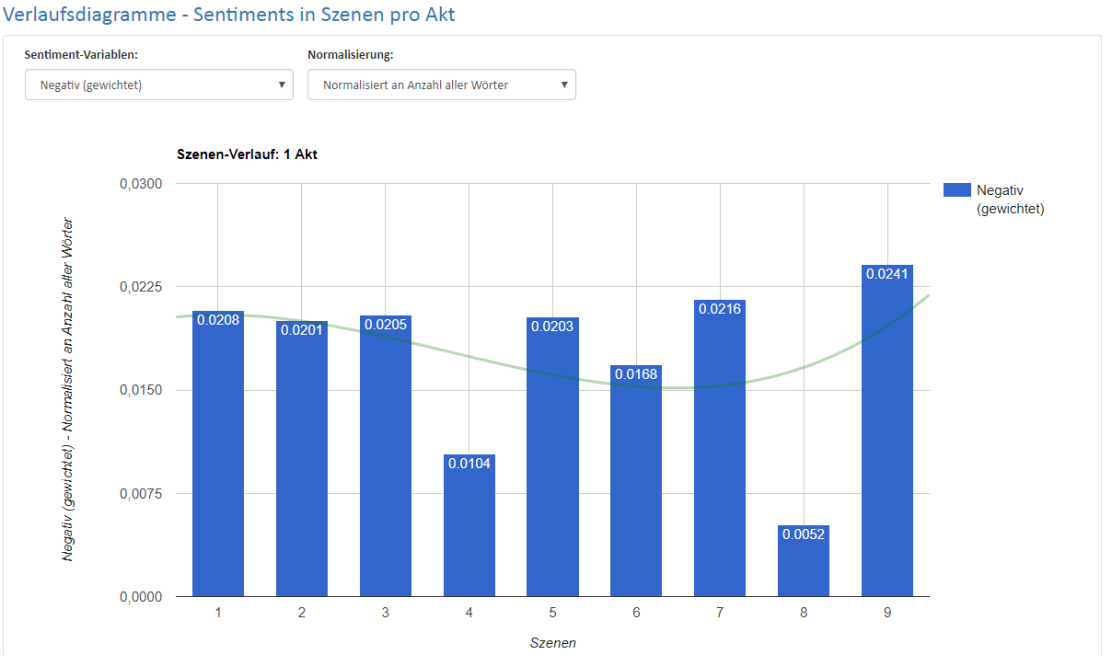


Abbildung 44: Visualisierung – Ausschnitt Verlaufsdiagramm – Sentiments in Szenen pro Akt

Durch die Normalisierung kann man die Szenen untereinander vergleichen und somit erkennen, dass sowohl Szene 4 als auch Szene 8 aus der sonst gleichmäßigen Sentiment-Ausprägung bezüglich Negativität abweichen und deutlich weniger negativ im Vergleich sind.

9.4.4.6 Kreisdiagramm – Sentiment-Anteile pro Szene

Auch das Kreisdiagramm für die Sentiment-Anteile pro Szene ist analog zu den bisherigen Anteilsdiagrammen aufgebaut, was die Wahl der Metriken und Normalisierun-

gen betrifft. Als einziger Unterschied kann man hier über Dropdown-Menüs die explizite Szene durch Auswahl des Aktes und der dazugehörigen Szenenummer auswählen. Das Kreisdiagramm passt sich dann an die neue Auswahl an. Wechselt man den Akt, dann wird die Auswahl standardmäßig auf die erste Szene zurückgesetzt.

Als Beispiel wird hier die Polaritätsverteilung aller Polaritätswörter für obige angesprochene Szene 8 des 1. Aktes angezeigt:

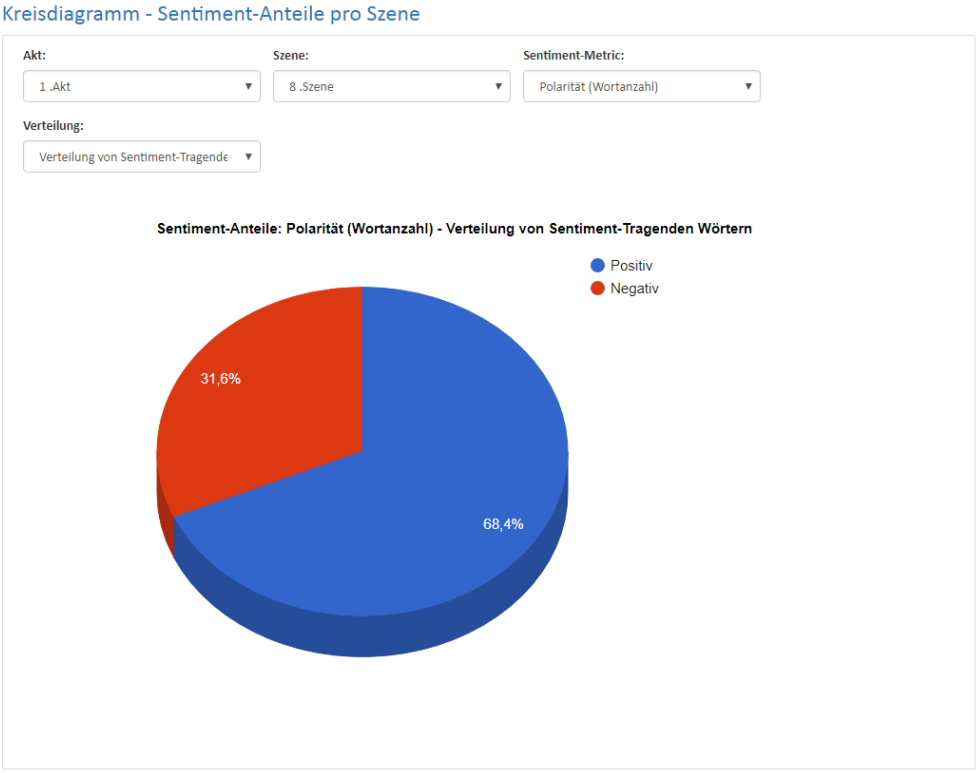


Abbildung 45: Visualisierung – Kreisdiagramm – Sentiment-Anteile pro Szene

Man kann in der Tat erkennen, dass deutlich mehr positive Wörter als negative Wörter in dieser Szene von Miss Sara Sampson enthalten sind, was konsistent zu der geringeren Negativitätsausprägung ist, die zuvor festgestellt wurde. Gleichzeitig kann man über das Hovern mit der Maus feststellen, dass die Szene recht kurz ist. Es handelt sich um lediglich 6 negative und 13 positive Wörter.

9.4.4.7 Verlaufsdiagramm – Szenen im Dramenverlauf

Über das nachfolgende Verlaufsdiagramm können die Sentiment-Werte von Szenen im Gesamt-Dramenverlauf betrachtet werden. Die Visualisierung unterscheidet sich von derjenigen aus Kapitel 9.4.4.5 dadurch, dass nicht zwischen Akten unterschieden wird und also keine separate Grafik pro Akt erstellt wird, sondern die Werte für die einzelnen Szenen ohne Unterbrechung angezeigt werden. Als Visualisierungsform wurde

hierbei ein Liniendiagramm anstatt eines Balkendiagramms gewählt, da dieses sich in der Darstellung als vorteilhafter erwies. Auf der x-Achse wird die Szene mit Akt abgetragen, auf der y-Achse der Sentiment-Wert. Als Beispiel wird anbei der Verlauf der Angst-Anteile an den Emotions-SBWs aufgezeigt:

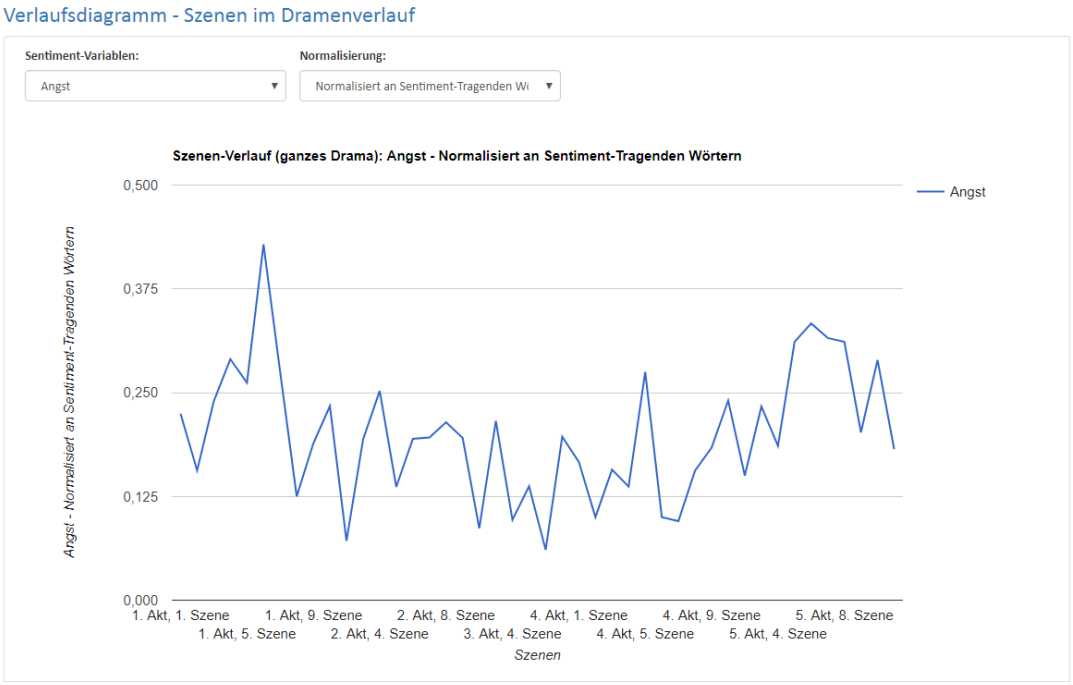


Abbildung 46: Visualisierung – Verlaufsdiagramm – Szenen im Dramenverlauf

Durch die Normalisierung kann man über diese Grafik erkennen wie der Anteil an Angst-Wörtern an Emotionswörtern sich im Dramenverlauf verhält. Er schwankt zwischen 10 und 40%. Indem man mit der Maus über die einzelnen Linien fährt können die expliziten Szenen-Informationen mit den exakten Werten eingesehen werden. Obiger Ausschlag beim 1. Akt, 6. Szene weist einen Angst-Anteil von 43% auf.

Bei Dramen mit zahlreichen Szenen kann man auf diese Weise einen guten Gesamtüberblick über den Sentiment-Verlauf erlangen und auch diejenigen Szenen mit besonderen Ausschlägen identifizieren. Auch hier sollte man die Normalisierungsoptionen beachten wenn man Szenen längenunabhängig betrachten will.

9.4.4.8 Verlaufsdiagramm – Repliken

Das finale Verlaufsdiagramm der strukturellen Analyse befasst sich mit der untersten Dramenebene dieses Projekts, der Repliken-Ebene. Analog zum Verlaufsdiagramm des letzten Kapitels kann man hier den Sentiment-Verlauf im Detail auf der Repliken-

Ebene betrachten. Es wird der explizite Wert pro Replik im Drama in Form eines Liniendiagramms angezeigt. Es wird nicht dem Vorschlag von Nalisnick und Baird (2013) in ihren Visualisierungen gefolgt, den summierten Wert anzugeben, also den Sentiment-Wert für jede Replik als akkumulierte Summe der vorherigen Repliken aufzufassen. Auf die hier vorliegende Weise können Repliken eher im Detail betrachtet werden, grundsätzliche Verläufe und Ausschläge identifiziert werden. Zukünftige Projekte können jedoch den Nutzen akkumulierter Werte bei der literaturwissenschaftlichen Interpretation untersuchen.

Die Auswahl der Metriken und Normalisierungen verläuft wie bei jedem bisherigen Verlaufsdiagramm. Auf der x-Achse wird als Verlaufsachse jede einzelne Replik angegeben. Auch hier wird eine Trendlinie eingebaut. Hovert man über die einzelnen Linienbestandteile wird die strukturelle Information für jede Replik (also Akt, Szene) sowie der genaue Sentiment-Wert angegeben. Folgende Grafik zeigt die Replikenwerte des Dramas Miss Sara Sampson für die Metrik Polarität (gewichtet) und absolut betrachtet:

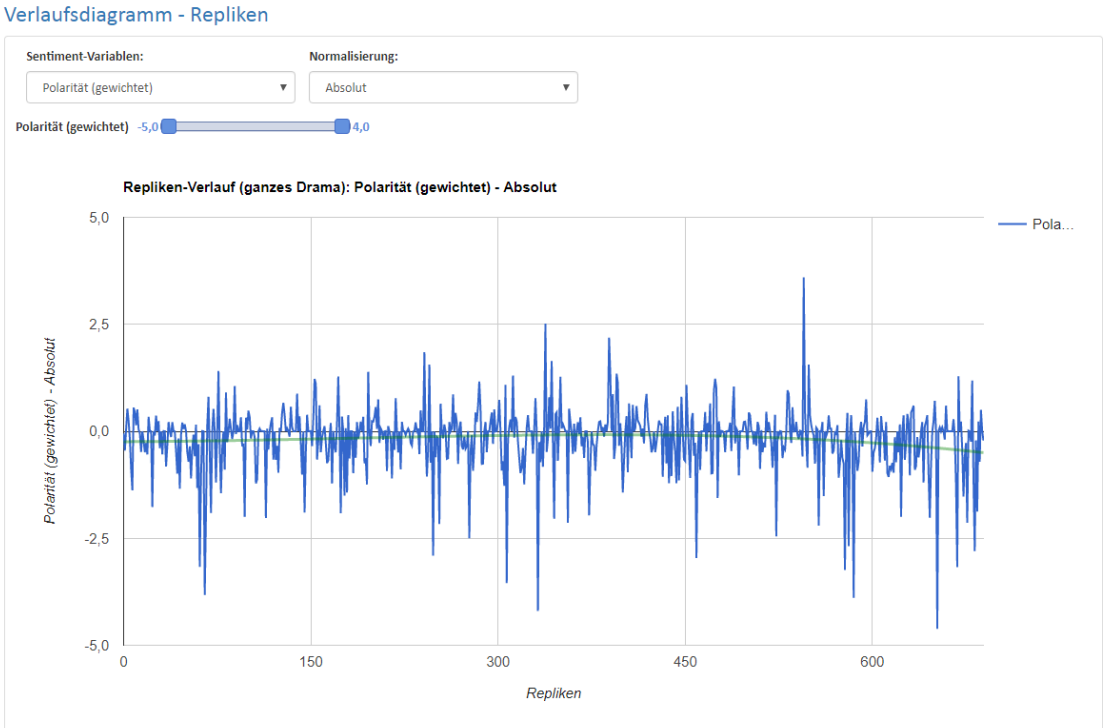


Abbildung 47: Visualisierung – Verlaufsdiagramm – Repliken

Über das Hovern über die Ausschläge kann man exakt feststellen welches besonders Polaritätsstarke Repliken sind. Man kann über die Trendlinie einen leichten Abfall gegen Ende des Dramas und damit eine Steigerung der Negativität feststellen. Auf Rep-

likenebene ist zu beachten, dass viele Repliken sehr kleine oder Null-Werte haben, da das Drama aus zahlreichen sehr kurzen Repliken besteht.

Für dieses Verlaufsdiagramm liegt zusätzliche Funktionalität vor. Man kann mit der Maus im Diagramm einen Bereich markieren, um hinein zu zoomen und sich diesen Bereich genauer anzuschauen. Hier wurde beispielsweise in den vierten Akt zum auffällig positiven Ausschlag gezoomt:

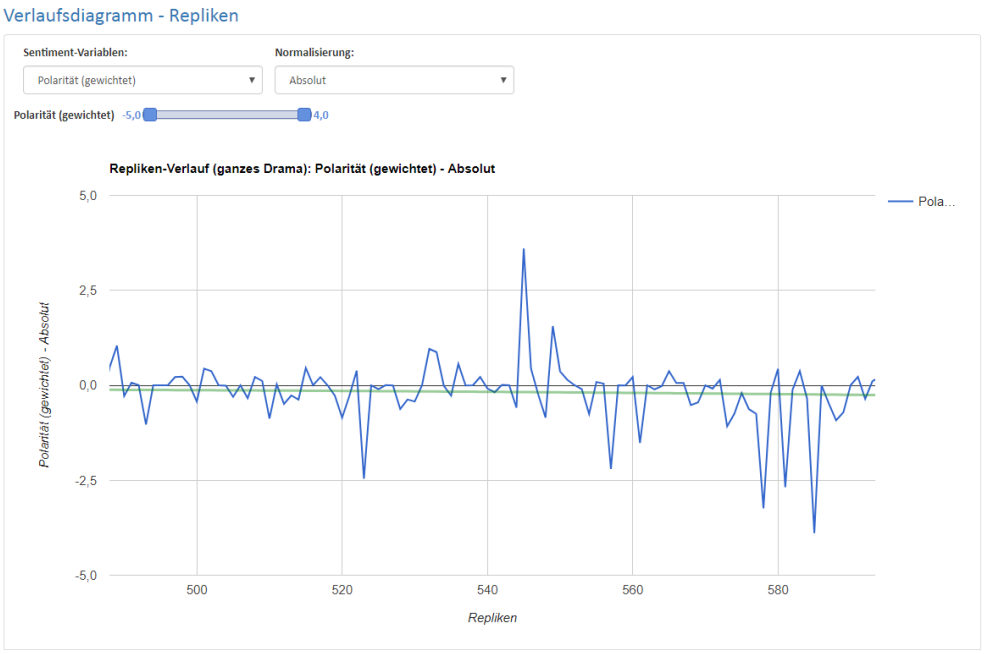


Abbildung 48: Visualisierung – Verlaufsdiagramm – Repliken (mit Zoom-Funktion)

Über einen Rechtsklick verlässt man die Zoom-Ansicht und kehrt wieder zur ursprünglichen Ansicht zurück. Über einen Slider kann man die Darstellung der Grafik auf einen speziellen Wertebereich einschränken. Dies ist vor allem bei den Emotions-Metriken, der Positivität und Negativität nützlich, um Repliken mit 0-Werten herauszufiltern und so nur relevante Metriken in der Gesamtgrafik einzusehen.

9.4.5 Sprecher-Analyse

Über den Unterpunkt Sprecher-Analyse können sprecherspezifische Sentiment-Verteilungen und Verläufe analysiert werden. Zur Kalkulation werden dazu die jeweiligen Repliken eines Sprechers genutzt. Die jeweiligen Sentiments beziehen sich also nicht direkt auf den Sprecher sondern sind Ausdruck, der von ihm verwendeten Sprache. Die Gliederung der Funktionalität für die Sprecher-Analyse ist grundsätzlich ähnlich zur strukturellen Analyse, insofern, dass man mit der obersten Ebene (Drama) im

UI anfängt und im untersten Gliederungspunkt bis auf die niedrigste Ebene, die Repliken-Ebene, geht. Ferner wurde ein spezieller Sprechervergleich eingebaut.

Für die folgenden Grafiken wurde das Drama Nathan der Weise gewählt. Dieses Drama hat ein verhältnismäßig großes Figuren-Repertoire für das Korpus.

9.4.5.1 Sentiments im ganzen Drama

Über diesen UI-Bereich kann man die Sprecher-Sentiments eines einzelnen Sprechers im Gesamt-Drama betrachten. Die Auswahl der Metriken und Normalisierungen ist analog zum gleichen Gliederungspunkt bei der strukturellen Analyse. Es können die Polarität (gewichtet) und die Polarität (Wortanzahl) betrachtet werden, die dann über drei Balken aufgeteilt nach Positivität, Negativität und Polarität visualisiert werden. Dazu können auf ähnliche Weise die Emotionen angezeigt werden. Über ein Drop-Down-Menü kann ein beliebiger Sprecher des Dramas ausgewählt werden. Über die Normalisierungsfunktion kann man wieder die Werte der Sprecher vergleichbar machen. Da jedoch stets nur ein Sprecher im Drama betrachtet werden kann, wird die weiter unten beschriebene Sprechervergleich-Visualisierung für gegenüberstellende Analysen empfohlen (siehe Kapitel 9.4.5.6).

Folgendes Beispiel gibt die absolute Zahl an positiven und negativen Wörtern für die Figur NATHAN an:

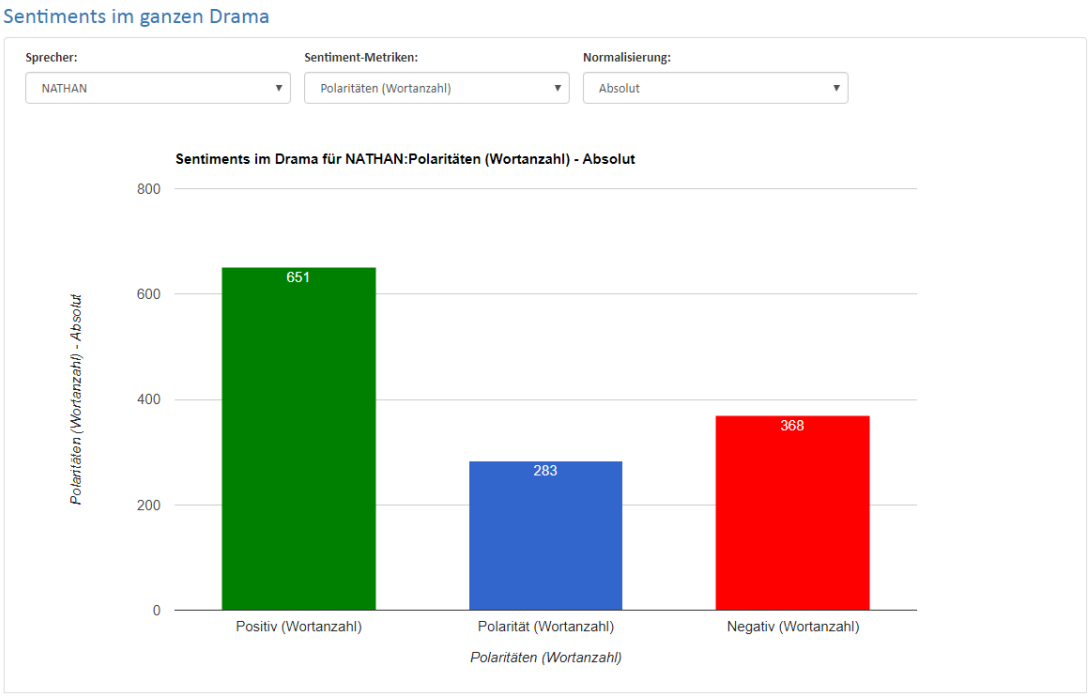


Abbildung 49: Visualisierung – Sentiments im ganzen Drama (Sprecher)

Man erkennt ein Übergewicht an positiven Wörtern.

9.4.5.2 *Verlaufsdiagramm – Sprecher-Sentiments pro Akt*

Über interaktive Diagramme lassen sich die dynamischen Sprecher-Sentiments im Verlauf des Dramas analysieren. Als erstes auf der Ebene des Aktes über ein Balkendiagramm. Wie üblich für Verlaufsdiagramme lassen sich alle Polaritäts- und Emotions-Metriken sowie der Normalisierungsfaktor über ein Drop-Down-Menü auswählen. Über Checkboxen kann man die zu analysierenden Sprecher auswählen. Es ist möglich einen Sprecher einzeln im Aktverlauf zu betrachten oder mehrere Sprecher zu analysieren und zu vergleichen. Wählt man mehrere Sprecher aus, werden gruppierte Balkendiagramme mit farblich untereschеidbaren Balken pro Akt pro ausgewählter Figur angezeigt.

Folgende Grafik visualisiert und vergleicht die Zornanteile an allen Emotions-SBWs im Dramenverlauf pro Akt:

Abbildung 50: Visualisierung – Verlaufsdiagramm – Sprecher-Sentiments pro Akt

Man erkennt bezüglich des UIs, dass man beliebig viele Sprecher zur Analyse und zum Vergleich auswählen kann. Tritt ein Sprecher in einem Akt nicht auf, wird dies über den Text kein Auftritt vermerkt. Über den Balken werden die gerundeten Werte ange-

zeigt. Werden mehr als drei Figuren über die Checkboxen ausgewählt, so werden die Werte aus Platzmangel nicht über den Balken angezeigt. Hovert man über die Balken können jedoch, wie bislang üblich, die exakten Werte eingesehen werden. Eine Trendlinie zeigt den Gesamtverlauf aller Figuren auf. Es ist verständlich, dass bei einer übermäßigen Auswahl vieler Figuren das Diagramm weniger übersichtlich wirkt. Potentiell ist dies zwar möglich, jedoch wird empfohlen die Auswahl im angemessenen Bereich zu halten oder für statische Vergleiche und Analysen, den noch weiter unten beschriebenen Sprechervergleich oder das Kreisdiagramm für Sentiment-Anteile zu verwenden.

Im vorliegenden Beispiel kann man über diese Visualisierung erkennen, dass sich die Zorn-Anteile an allen Emotionen der beiden Figuren vor allem in Akt 3 und Akt 4 merklich unterscheiden, während sie ansonsten sehr ähnlich sind. Je nach Analyseziel empfiehlt es sich wieder statt der absoluten normalisierte Werte zu wählen, um Figuren bezüglich des Redeanteils vergleichbarer zu machen.

9.4.5.3 Verlaufsdiagramm – Sprecher-Sentiments in Szenen pro Akt

Analog zur Visualisierung auf struktureller Ebene können Sprecher-Sentiments auch in Szenen pro Akt in Form von normalen und gruppierten Balkendiagrammen dargestellt werden. Wie bei den meisten Sprecher-Analysen können über Checkboxen, die zu analysierenden Sprecher ausgewählt werden, also einer bis potentiell alle. Standardmäßig wird der erste Sprecher des Dramas angezeigt. Die Auswahl der Sentiment-Variablen und der Normalisierung umfasst die üblichen Maße und funktioniert analog zu allen Verlaufsgrafiken. Es wird für jeden Akt ein Balkendiagramm erstellt. Auf der jeweiligen x-Achse wird die Szenennummer abgetragen und auf der y-Achse die Sentiment-Metrik. Eine Trendlinie visualisiert den generellen Verlauf über alle angegebenen Sprecher hinweg. Fehlt eine Figur in einer Szene wird das wieder mit dem Hinweis „kein Auftritt“ angegeben, jedoch nur, wenn lediglich eine Figur ausgewählt wurde. Bei der Auswahl mehrere Figuren musste man feststellen, dass die Grafiken überladen und verwirrend wirken, wenn stets „kein Auftritt“ angegeben wird. Bei mehreren Figuren erkennt man ein Fehlen einfach dadurch, dass kein Sentiment-Wert für die Szene angezeigt wird. Ebenso hat es sich aus Übersichtsgründen als vorteilhaft erwiesen die gerundeten Sentiment-Werte nur anzugeben, wenn genau eine Figur ausgewählt wurde. Bei zwei und mehr ausgewählten Figuren kann dies sonst zu einem visuell

überladenen Eindruck führen und sich Werte z.B. überschneiden. Die Hover-Funktionen der Maus funktionieren wie bei den anderen Grafiken zur Detaileinsicht der genauen Werte jedoch wie bisher bekannt.

Als Beispielvisualisierung werden wieder die Figuren NATHAN und SALADIN gewählt. Diesmal mit der gewichteten Polarität normalisiert an der Anzahl der Wörter um eine direkte Vergleichbarkeit zu ermöglichen. Es wird Akt 3 betrachtet:

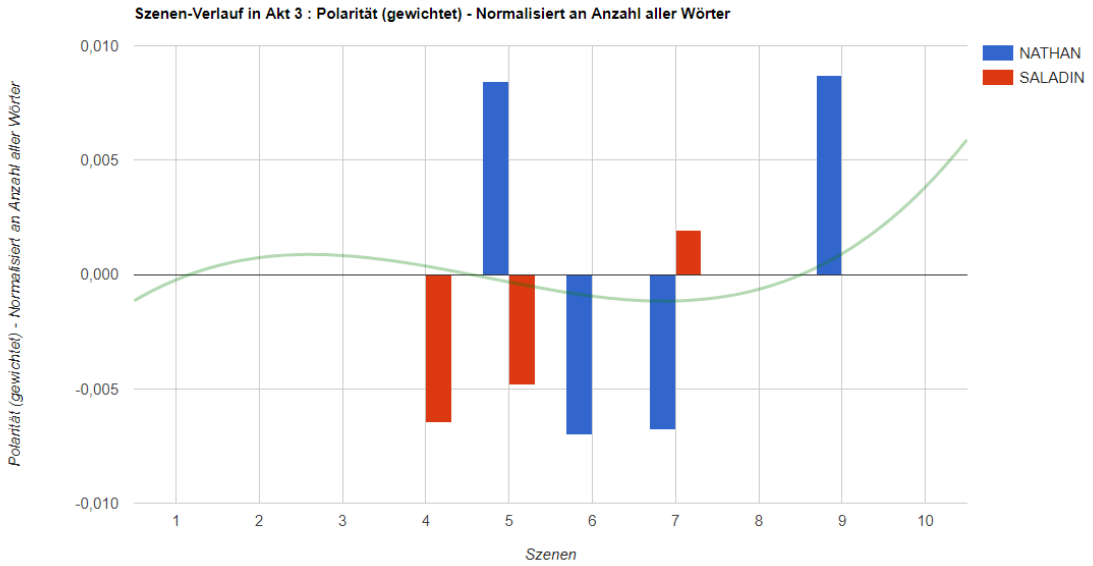


Abbildung 51: Visualisierung – Ausschnitt – Sprecher-Sentiments in Szenen pro Akt

Man erkennt, dass diese Grafiken für die vergleichende Nutzung eher angebracht sind wenn Figuren häufig in Szenen gleichzeitig auftreten. Im vorliegenden Fall treten Nathan und Saladin in Szene 5 und 7 auf. Man stellt stark unterschiedliche Polaritätsausprägungen fest, Nathan sehr viel positiver in Szene 5 und sehr viel negativer in Szene 7.

9.4.5.4 Verlaufsdiagramm – Szenen im Dramenverlauf pro Sprecher

Auch für die Sprecher-Analyse kann man den Szenenverlauf anstatt pro Akt wie oben, für den gesamten Dramenverlauf ohne Unterbrechung betrachten, um einen grundsätzlichen Gesamtüberblick für einen Sentiment-Metrik zu bekommen. Hierfür wurde analog zur strukturellen Analyse ein Liniendiagramm zur Visualisierung gewählt. Auf der x-Achse befinden sich wieder die Szenennummern und auf der y-Achse die entsprechend ausgewählten Werte für die Sentiment-Metrik pro Szene und Sprecher. Man kann wieder über Checkboxen eine beliebige Zahl Figuren auswählen. Jede Linie repräsentiert dabei eine Figur. Als Beispiel werden wieder diesmal die Figuren Nathan,

Daja, Recha und Saladin ausgewählt. Als Metrik wird exemplarisch die gewichtete Polarität als absoluter nicht-normalisierter Wert selektiert:

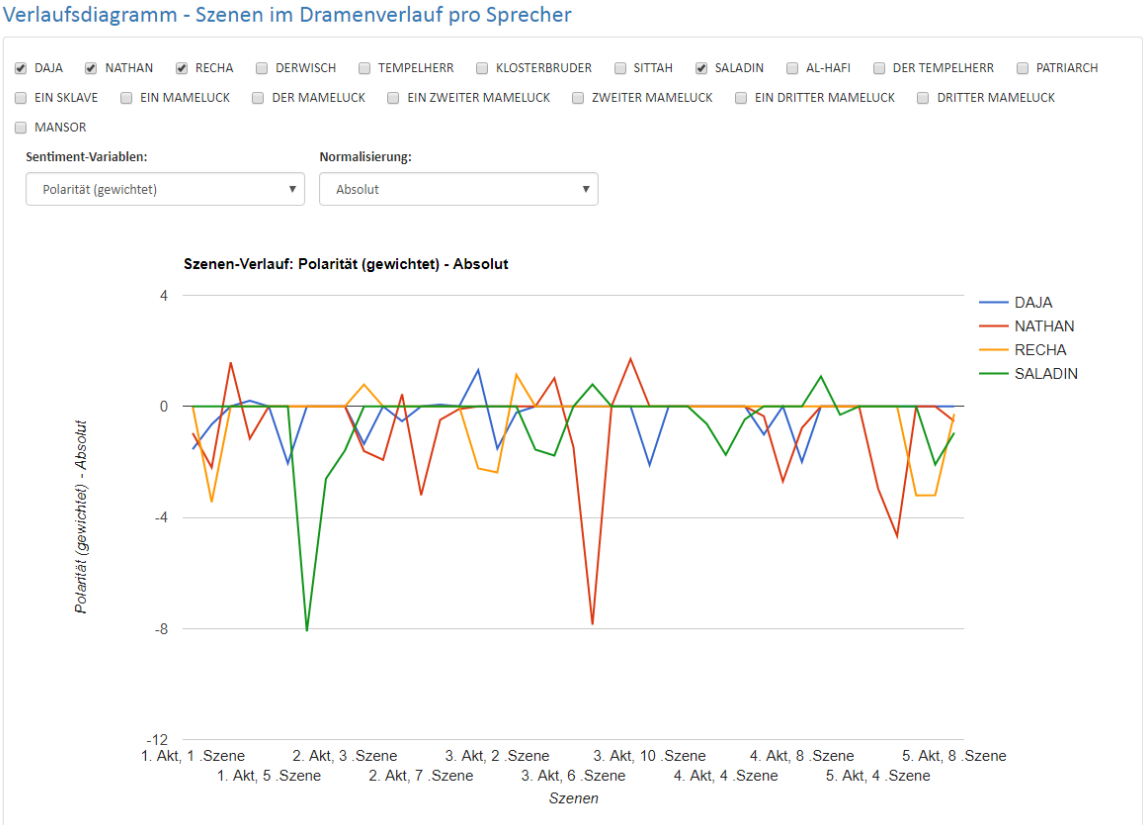


Abbildung 52: Visualisierung – Verlaufsdiagramm – Szenen im Dramenverlauf pro Sprecher

Man erkennt für einige Figuren besondere Zusammenhänge. Indem man mit der Maus über die einzelnen Linien fährt, kann man über Tooltips Detailinformationen zur Analyse erhalten. Für Saladin entdeckt man eine besonders negativ konnotierte Szene im 2.Akt, 1. Szene, für Nathan eine im 3.Akt, 7. Szene wobei in selbiger Szene Saladin wieder eher positive Sentiments anzeigt. Man kann insgesamt für alle Figuren eine Tendenz zu negativen Ausschlägen identifizieren. Auch hier erhalten Figuren die in Szenen nicht auftreten automatisch einen Null-Wert. Bezüglich dieses speziellen Diagramms ist zu beachten, dass die absoluten Werte betrachtet werden, das heißt größere Ausschläge können auch dadurch entstehen, dass einzelne Figuren mehr Sprechanteile haben und somit größere Polaritäten generieren. Zur Vergleichbarkeit von Sprechern kann man jedoch, wie bisher häufig angesprochen, Normalisierungsoptionen auswählen. Die Auswahl absoluter Werte kann je nach Fragestellung jedoch auch sinnvoll erscheinen wenn man grundsätzlich Sentiment-starke Einheiten finden will, unabhängig von der direkten Vergleichbarkeit. Eine lange Szene voller starker Negativität kann als

insgesamt negativer betrachtet werden als eine sehr kurze Szene mit dichter Negativität.

Es sei auch hier anzumerken, dass bei der Auswahl von zu vielen Sprechern die Übersichtlichkeit nicht mehr eingehalten werden kann. Es bleibt dem Nutzer aber selbst überlassen wie viele Figuren er tatsächlich betrachten will.

9.4.5.5 Verlaufsdiagramm pro Sprecher – Repliken

Bezüglich des Dramenverlaufs kann man Sprecher nun abschließend auch auf der kleinsten vorhandenen Dramenebene der Repliken-Ebene einzeln oder vergleichend analysieren. Das Liniendiagramm funktioniert dabei analog zur äquivalenten Version bei der strukturellen Analyse. Einziger Unterschied ist, dass man über Checkboxen verschiedene Sprecher mit ihren Werten zur Anzeige auswählen kann. Die sonstige Auswahl über Drop-Down-Menüs ist analog zum Bisherigen. Über die Repliken-Ansicht kann man sich Sentiment-Verläufe im Gesamtdrama im Detail aber auch vergleichend anschauen. In folgendem Beispiel wird die gewichtete Polarität normalisiert an der Anzahl der Wörter für die drei Figuren AL-HAFI, NATHAN und SALADIN pro Replik angezeigt:

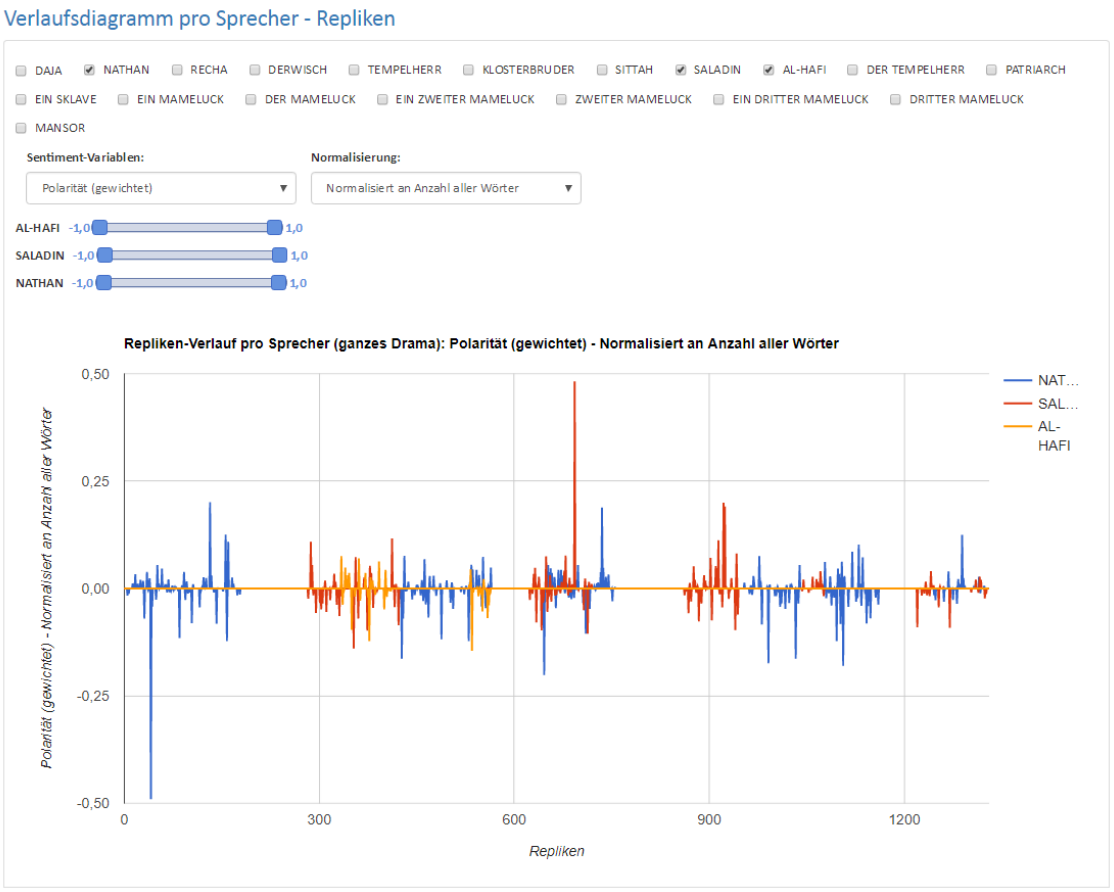


Abbildung 53: Visualisierung – Verlaufsdiagramm pro Sprecher – Repliken

Man kann vereinzelte deutliche Ausschläge für alle drei Figuren erkennen. Für die Figur Saladin lassen sich positive Cluster identifizieren. Die Figur Al-Hafi weist keine besonderen Ausschläge auf. Indem man mit der Maus über Linien hovert, wird exakt angezeigt, um welche Replik es sich handelt und welchen Wert diese genau einnimmt. In der Grafik sieht man auch, dass alle Sprecher für Repliken an denen sie nicht beteiligt sind, den Wert 0 zugewiesen bekommen, was dazu führt, dass eine Linie auf der 0-Achse stets sichtbar ist. Diese Lösung war notwendig, um die Visualisierung ansprechend zu gestalten, da das Liniendiagramm sonst aus großen häufigen Lücken und Punkten bestehen würde. Hovert man jedoch über eine derartige Replik wird über einen Tooltip die korrekte Information angezeigt, dass diese Replik nicht zu dem jeweiligen zur Linie gehörenden Sprecher gehört. Stattdessen wird der korrekte Sprecher mit allen weiteren Repliken-Informationen angezeigt.

Auch hier wird wieder die Funktionalität geboten, über Markieren eines Bereiches in den markierten Bereich hinein zu zoomen. Auf diese Weise kann man besonders auffällige Abschnitte genauer einsehen. Als weiter Funktionalität kann man über Slider

wieder den Wertebereich eingrenzen der pro Figur betrachtet werden soll. Diese Funktion kann für einige Metriken als nützlich angesehen werden, um durch das Entfernen von Repliken mit sehr geringen Werten die Übersichtlichkeit zu steigern, ist jedoch für vereinzelte Metrik (Polaritäten, normalisierte Werte) weniger brauchbar. Da es sich um ein Expertentool handelt hat man sich jedoch dazu entschieden die Funktion auch bei der Sprecher-Analyse beizubehalten. Man sollte des Weiteren auch hier beachten, dass die Übersichtlichkeit der Grafik unter übermäßiger Auswahl von Figuren aufgrund von Informationsüberladung stark eingeschränkt sein kann, potentiell jedoch möglich ist.

9.4.5.6 Statischer Sprechervergleich

Unter dem statischen Sprechervergleich wird nun ein UI-Abschnitt präsentiert mit dem Sprecher für die üblichen Verlaufs-Metriken für jede beliebige strukturelle Einheit über ein Balkendiagramm miteinander direkt verglichen werden können. Grundsätzlich sind derartige Vergleiche auch mit obigen Diagrammen möglich. Dafür müssen jedoch alle Sprecher ausgewählt werden, worunter wiederum die Übersicht stark leidet. Um die Visualisierung des Vergleichs zu verbessern, wurde die Diagrammerstellung über den statischen Sprechervergleich implementiert.

Über Drop-Down-Menüs wählt man nun neben der Metrik und Normalisierung auch die exakte strukturelle Einheit zur Analyse aus. Man kann das Gesamt-Drama, jeden einzelnen Akt, aber auch einzelne Szenen eines Aktes betrachtet. Für das ganze Drama wählt man mit dem Drop-Down Drama-Akt den Punkt Gesamt aus. Das andere strukturelle Drop-Down-Menü passt sich dann hierfür an. Will man einen Akt ansehen wählt man über Drama-Akt den Akt und über Akt-Szene gesamt. Gesamt für Akt-Szene wird automatisch bei jedem Akt-Wechsel gewählt. Möchte man eine spezielle Szene analysieren, kann man diese über Akt-Szene auswählen.

In einem Balkendiagramm werden dann die Werte aller Sprecher, die in der entsprechenden strukturellen Einheit erscheinen angezeigt. Die Anzeige der Sprecher ist dabei nach Höhe des Sentiment-Wertes geordnet, so dass man schnell identifizieren kann welcher Sprecher das Maximum für eine Metrik in einer Einheit annimmt.
Folgendes Beispiel zeigt die geordneten Werte für den negativen Wortanteil an allen Polaritäts-Wörtern (also normalisiert an Sentiment-Tragenden Wörtern) für den 3. Akt von Nathan der Weise:

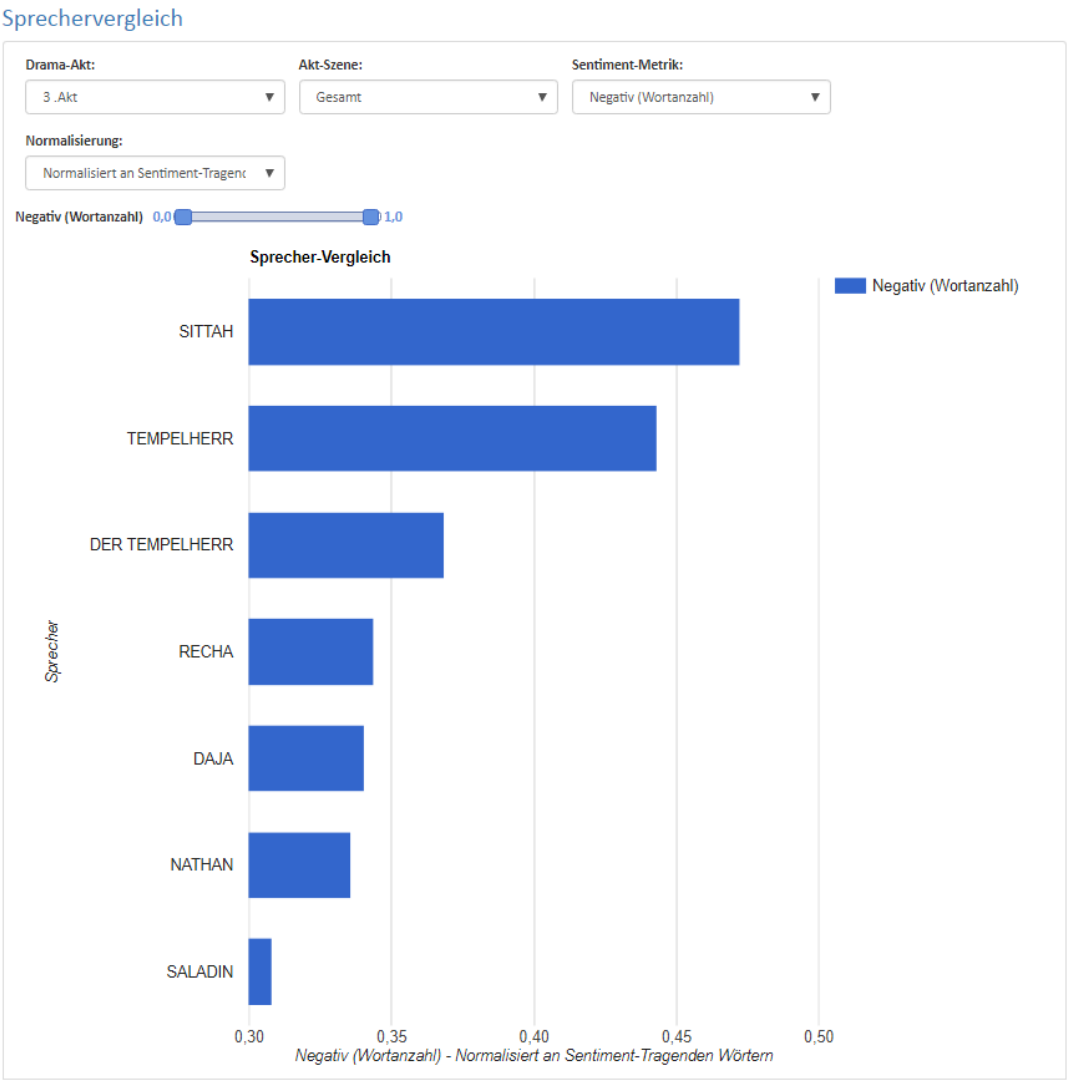


Abbildung 54: Visualisierung - Sprechervergleich

Man erkennt, dass in dem Akt 7 Figuren auftreten. SITTAH hat den höchsten Anteil an negativen Wörtern in seinen Sprechakten, nämlich etwa 47%, wie man durch hovern über den Balken im Detail erfahren kann. Saladin hingegen hat einen auffällig geringen Anteil an negativen Wörtern von lediglich 31%. Für diesen Akt kann man also auf Basis der SA sagen, dass Sittah die negativsten Redeanteile hat, während die von Saladin von geringer Negativität geprägt sind.

Auch hier kann man wieder über einen Slider den zu betrachtenden Wertebereich einschränken. Dies ist wieder nur bei einigen Metriken eine nützliche Funktion, um die Zahl der relevanten Sprecher zu vermindern und sich auf die Sprecher mit den höchsten Werten zu fokussieren.

9.4.5.7 Kreisdiagramm - Sentiment-Anteile von Sprecher

Auch für die Sprecher kann man nun Sentiment-Anteile über ein Kreisdiagramm einsehen. Die Funktionalität ist dabei an den bekannten Kreisdiagrammen aus der strukturellen Analyse orientiert. Als Sentiment-Metrik kann wieder die Verteilung der Polaritäten, der Emotionen sowie der Emotion vorhanden betrachtet werden. Entweder in Bezug auf die zugehörigen Sentiment-Tragenden Wörter oder in Bezug auf alle Wörter. Für das Sprecher-Kreisdiagramm kann man nun jedoch die Verteilung für das Drama gesamt, einzelne Akte und einzelne Szenen anschauen. Für die Auswahl jeder beliebigen Ebene werden über das Sprecher-Drop-Down-Menü die Auswahlmöglichkeiten für Sprecher angepasst, je nachdem welche Sprecher in einer strukturellen Einheit vorkommen. Über die Auswahl Gesamt beim Drama-Drop-Down kann wieder das gesamte Drama eingesehen werden, ähnliches gilt für den Akt beim Akt-Szene-Drop-Down. Je nach Auswahl passen sich auch die strukturellen Drop-Downs an die jeweilige Vorauswahl an.

Als Beispiel wird diesmal die Sentiment-Metrik-Option „Emotion vorhanden“ gewählt mit der Figur SITTAH für den 3. Akt, 4. Szene. Als Verteilungsart wird die Verteilung von allen Wörtern selektiert. Eine Verteilung von Sentiment-Tragenden Wörtern ist für diese Sentiment-Metrik nicht sinnvoll, da diese immer 100% beträgt, da bei allen Emotionstragenden Wörtern trivialerweise eine Emotion vorhanden ist:

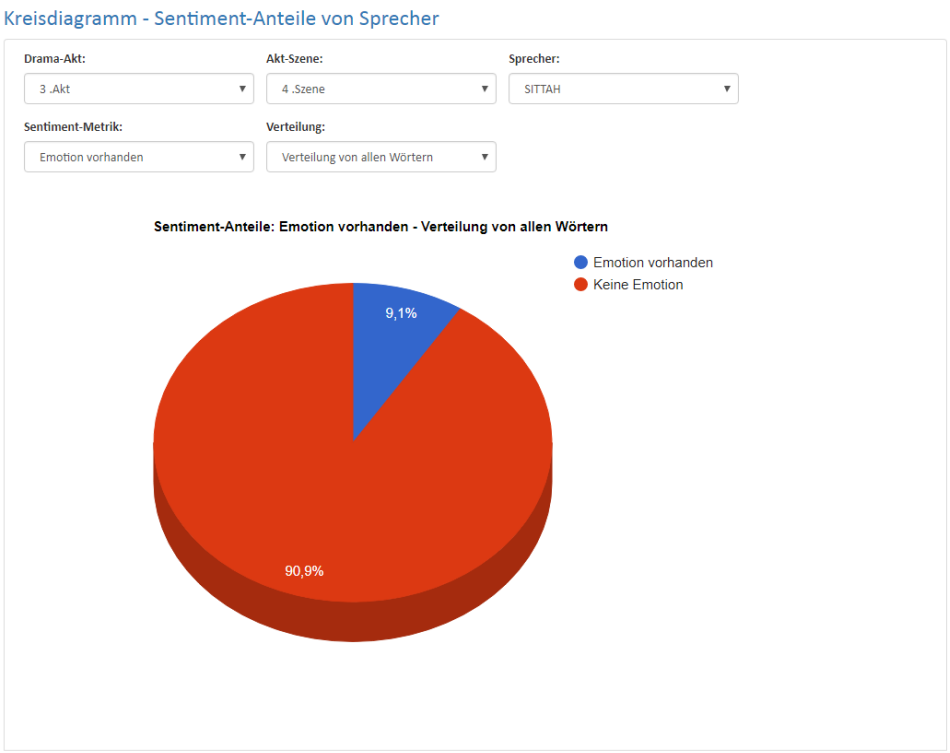


Abbildung 55: Visualisierung – Kreisdiagramm – Sentiment-Anteile von Sprecher

Man erkennt, dass nur wenige Wörter, etwa 9% mit Emotionen gemäß NRC-Lexikon konnotiert sind. Durch hovern auf die Bereich kann man die exakten Werte erfahren: lediglich 22 Wörter sind Emotionswörter gegenüber 220 nicht emotionstragenden Wörtern.

9.4.6 Sprecher-Beziehungen (Charakter-zu-Charakter-Sentiment)

Als letzter Menüpunkt können Nutzer Charakter-zu-Charakter-Sentiments als Verläufe und Verteilungen betrachten. Diese werden Sprecher-Beziehungen genannt, obschon weniger eine direkte Beziehung dargestellt wird als das isolierte Sentiment eines Ausgangssprechers auf einen Zielsprecher. Berechnet wird dies über die Repliken vom Ausgangsprecher, die nach Repliken des Zielsprechers erscheinen. Dieses Sentiment wird gemäß Nalisnick und Baird (2013) als Heuristik für die Sprecher-Beziehung betrachtet. Auch in diesem Menüpunkt lassen sich derartige Sprecherbeziehungen auf der Verlaufsebene und der Verteilungsebene betrachten.

Zur Betrachtung und Erklärung von Funktionen anhand von Beispielen wird für diesen Abschnitt das Drama „Der junge Gelehrte“ gewählt. Es hat eine überschaubare Zahl an Figuren, deren Hauptfiguren häufig miteinander in Dialoge treten. Somit ist das Drama gut für die Erläuterung der Sprecher-Beziehungs-Komponente geeignet.

9.4.6.1 Sprecher-Beziehungs-Sentiments im Drama

Konsistent zu den bisherigen anderen Menüpunkten kann man als erste Sprecher-Beziehungen auf der höchsten Ebene, der Dramenebene, in Form eines Balkendiagramms betrachten. Man kann die Sentiment-Ausprägung für die für Verlaufsgrafiken üblichen Metriken einsehen (obschon hier kein expliziter Verlauf angezeigt wird). Die Visualisierung verläuft grundsätzlich analog zu den bisherigen Verfahren. Auch die Auswahl der Metriken verläuft über äquivalente Drop-Down-Menüs. Es wird nun noch ein zusätzliches Drop-Down-Menü angeboten, um den Ausgangssprecher der Sentiment-Beziehung auszuwählen. Als Ziel-Sprecher werden alle Sprecher angezeigt zu denen gemäß Heuristik eine Beziehung kalkuliert werden kann, also wenn mindestens eine Replik vor einer Replik des Ausgangssprechers erscheint und somit heuristisch von einem aufeinander bezogenen Dialog ausgegangen wird. Über Checkboxes ist es möglich dann die zu betrachtenden Ziel-Sprecher auszuwählen. Die Checkboxes passen sich an den Ausgangssprecher an. Man kann beliebig viele Zielsprecher auswählen. Auf diese Weise kann man nicht nur einzelne Charakter-zu-Charakter-Sentiments betrachten, sondern diese auch aus Sicht des Ausgangssprechers vergleichen. Bezüglich der Vergleichbarkeit gelten die gleichen Besonderheiten wie sonst, so dass man je nach Anwendungsziel Normalisierungsoptionen auswählen sollte, um die Länge als Einflussfaktor zu relativieren. Wechselt man den Ausgangssprecher, passen sich je nach Vorhandensein einer Beziehung auch die Sprecher an.

Auf der x-Achse werden die jeweiligen Zielsprecher angezeigt, auf der y-Achse die Werte für die Sentiment-Metrik als gerundete Zahlen über den Balken. Genauere Daten kann man hier, wie in den anderen Grafiken auch, durch hovern über die Balken einsehen. Folgendes Beispiel zeigt die Sentiment-Beziehungen mit der Figur Damis als Ausgangssprecher und die Polarität (gewichtet) normalisiert an der jeweiligen Anzahl der Wörter jeder Beziehung:

Abbildung 56: Visualisierung – Sprecher-Beziehungs-Sentiments im Drama

Es wurden nur die relevanten Figuren ausgewählt. Die Figur Damis hat gemäß Kalkulation auch eine Sentiment-Beziehung zu sich selbst. Dies kann gemäß Algorithmus durchaus vorkommen, wenn die gleiche Figur mehrmals hintereinander einen Sprechakt tätigt. Auch findet keine szenenweise Zurücksetzung statt, das heißt Beziehungen können über szenenübergreifende Repliken-Abfolgen entstehen. Aus den genannten Gründen kann eine Figur eine Beziehung mit sich selbst haben. Es bleibt dem Experten-Nutzer des Tools überlassen wie er derartiges interpretiert. An dieser Stelle wurde die Selbstbeziehung über die Chechbox ausgelassen.

Man erkennt, dass Damis' Meinungen und Sentiments zu anderen Figuren gemäß dieser Heuristik meist negativ sind, abgesehen von der Figur der JULIANE. Zur Figur des ANTON besteht die negativste Beziehung.

9.4.6.2 Sprecher-Beziehungs-Sentiments pro Akt

Au der Verlaufsebene lassen sich Sprecher-Beziehungen pro Akt und pro Szene (siehe unten) betrachten. Bei der Verlaufsebene pro Akt handelt es sich um ein Balkendiagramm bzw. gruppiertes Balkendiagramm bei der Auswahl mehrerer Zielpersonen. Auf der y-Achse werden die einzelnen Akte abgetragen, die y-Achse gibt wieder die gewählte Sentiment-Metrik wider. Es können die für die Verlaufsebene üblichen Met-

riken, wieder über ein Drop-Down-Menü gewählt werden. Genau wie bereits bei obigen Diagramm können über Checkboxes beliebig viele oder wenige Ziel-Sprecher des Ausgangssprechers für die Visualisierung der Sentiment-Beziehungen ausgewählt werden. Ist mehr als eine Figur ausgewählt, werden die Balken pro Akt gruppiert und die einzelnen Zielfiguren der Beziehung sind über Farben unterscheidbar. Die genaue Funktionalität wird anhand der Figur Lisette und ihren Sentiment-Beziehungen mit der Variable Polarität (gewichtet), wieder normalisiert an der Anzahl der Wörter:

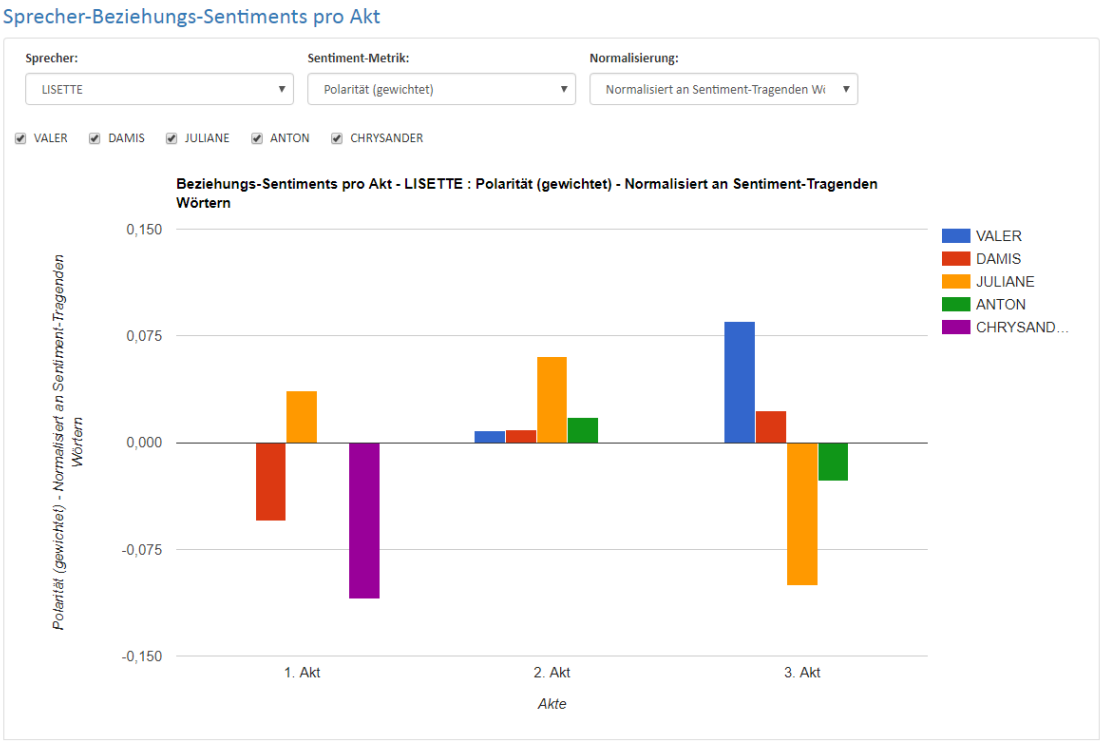


Abbildung 57: Visualisierung – Sprecher-Beziehungs-Sentiments pro Akt

Man erkennt Entwicklungen mit einigen Figuren über den Dramenverlauf. Die negativste Ausprägung lässt sich für die Figur CHRYSANDER im 1. Akt konstatieren. Die Meinung bzw. das Sentiment zu DAMIS ist im ersten Akt auch eher negativ, wird dann aber positiv mit Fortschreiten des Dramas. Für die Beziehung zur Figur der JULIANE kann man das Gegenteil verzeichnen. Zu dieser Figur ist das Sentiment erst positiv ,um im 3. Akt einen starken Ausschlag in die Negativität aufzuzeigen. Die exakten Werte kann man wie üblich über hovern über die Balken einsehen.

9.4.6.3 Sprecher-Beziehungs-Sentiments pro Szene

Analog zu den Sprecher-Beziehungs-Sentiments pro Akt kann man nun auch die Beziehungs-Sentiments pro Szene einsehen. Im Gegensatz zu den anderen Komponenten

des Tools, also die strukturelle Analyse und die Sprecher-Analyse, wird hier jedoch kein Balkendiagramm pro Akt präsentiert. Die Visualisierung dieser Art hat sich aufgrund der selten vorhandenen Sentiment-Beziehungen pro Szene als insgesamt wenig aussagekräftig und bei der Wahl mehrerer Ziel-Sprecher als sehr unübersichtlich herausgestellt. Deswegen hat man sich hier auf ein Liniendiagramm beschränkt, bei der auf der x-Achse die Szenennummern abgetragen werden und auf der y-Achse die dazugehörigen Sentiment-Werte. Wieder über Checkboxes können beliebig viele oder wenige Zielsprecher zur Analyse und für den Vergleich gewählt werden. Jeder Zielsprecher wird als Linie visualisiert. Ein Zielsprecher kann nur einen Wert über 0 erhalten, wenn er wenigstens einmal in einer Szene mit einer Figur gemäß der gewählten Heuristik in einen Dialog tritt. Als Beispiel werden die Sentiment-Beziehungen der Figur ANTON gewählt mit allen Figuren außer zu sich selbst. Als Metrik wird die Polarität (Wortanzahl) gewählt. Durch die Normalisierung an der Anzahl der Wörter wird der Einfluss der Länge von Text pro Beziehung auf die Polaritätsstärke relativiert:

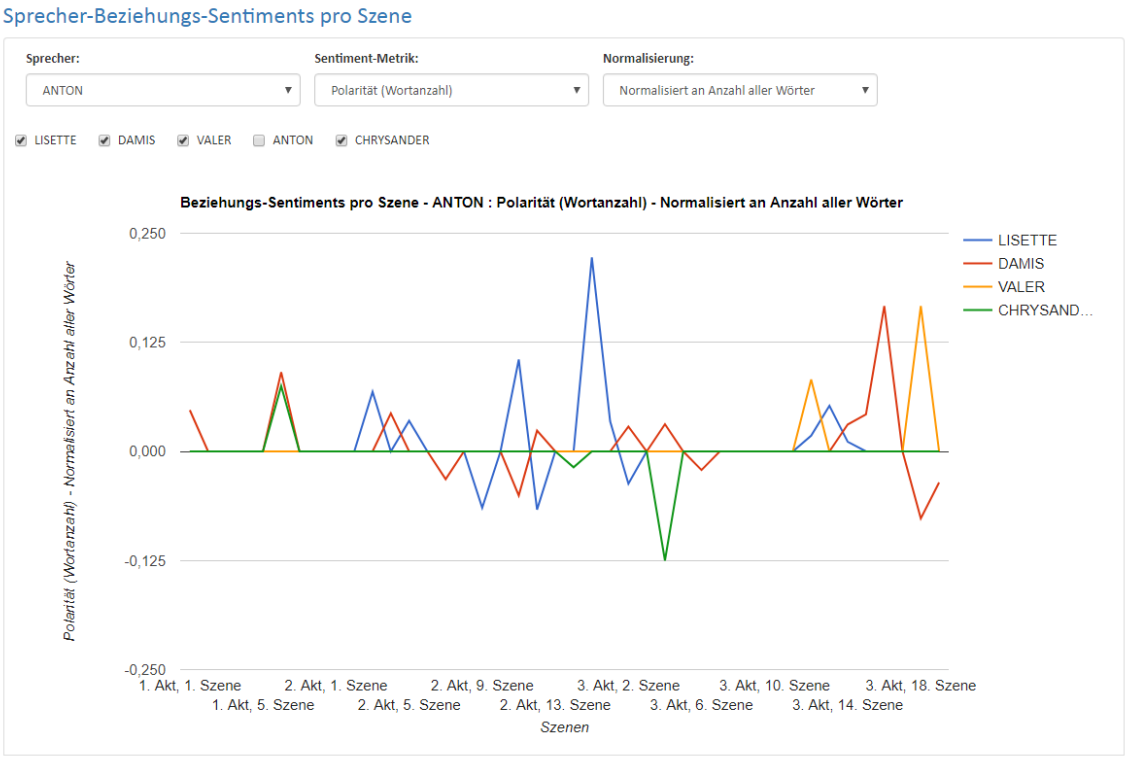


Abbildung 58: Visualisierung – Sprecher-Beziehungs-Sentiments pro Szene

Man kann verschieden Feststellungen machen. Auffällig sind starke Ausschläge für die Figur Lisette, vor allem 2.Akt, 15. Szene (Mitte der Grafik). Sowohl für DAMIS als auch VALER erkennt man gegen Ende des Dramas positive Extrema, die dann jedoch im

Fall von DAMIS in das Negative umschlagen. Indem man über die einzelnen Linienabschnitte hovert, können über einen Tooltip detaillierte Metrik-Angaben mit den exakten Werten eingesehen werden. Es wird deutlich, dass bei der Auswahl mehrerer Figuren die Übersichtlichkeit problematisch wird. Es wird deswegen empfohlen sich je nach Fragestellung auf eine angemessene Menge an Figuren zu beschränken oder Figuren in kleinen Gruppen anzuschauen.

9.4.6.4 Kreisdiagramm - Sentiment-Anteile von Sprecherbeziehungen

Als letzte Visualisierung werden nun für Sprecherbeziehungen auch Sentiment-Anteile in einem Kreisdiagramm illustriert. Auf diese Weise kann man sich für alle Sentiment-Beziehungen auf jeder strukturellen Ebene, insofern die Beziehung zu einer Figur gemäß Heuristik vorhanden ist, Sentiment-Verteilungen bezüglich Polarität und Emotionen ansehen. Die Drop-Down-Auswahl für die Sentiment-Metrik und die Normalisierung verläuft analog zu den bisherigen Kreisdiagrammen zur Visualisierung von Sentiment-Anteilen aus den anderen Tool-Komponenten. Wie bei der Auswahl der Sprecher-Analyse kann man auch hier über Drop-Down entscheiden, ob man Beziehungen der Dramen-, Akt- oder Szenen-Ebene auswählen möchte. Dementsprechend kann man sich die Verteilungen zu Beziehungen im ganzen Drama, in einem Akt oder in einer Szene ansehen. Je nach ausgewählter struktureller Einheit kann man über ein weiteres Drop-Down-Menü aus allen für diese Einheit verfügbaren Ausgangs-Sprecher auswählen. Über ein Drop-Down, das sich an die Auswahl der vorherigen Menüs anpasst, ist es sodann möglich den Ziel-Sprecher aus allen zum Ausgangssprecher gehörenden Ziel-Sprechern auszuwählen. Die Funktionalität des Kreisdiagramms ist ansonsten wie üblich, indem man mit der Maus über eine Kreiskategorie fährt, kann man die detaillierten Metrik-Angaben betrachten, also die Kategorie, den genauen Wort und die exakte Wortanzahl. Folgendes Beispiel zeigt die Emotionsverteilung für die Figur LISETTE ausgerichtet auf die Figur CHRYSANDER für das gesamte Drama:

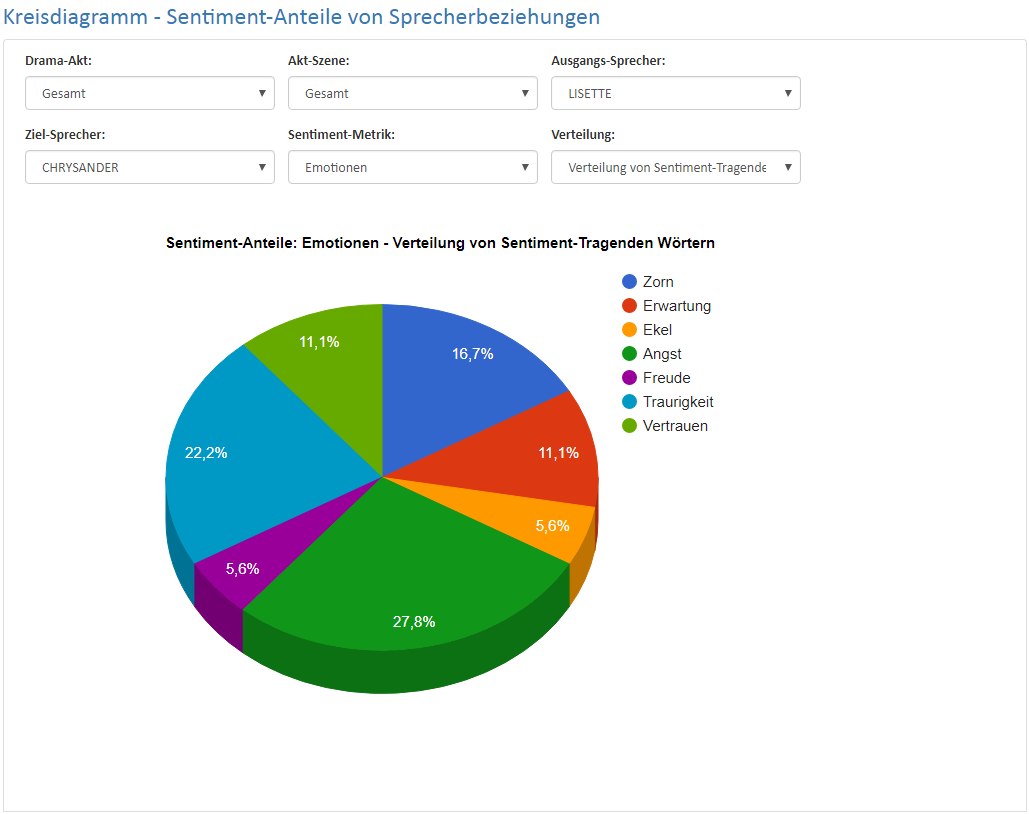


Abbildung 59: Kreisdiagramm – Sentiment-Anteile von Sprecherbeziehungen

Man kann erkennen, dass negative Emotionen mit einem höheren Anteil vertreten sind, vor allem Angst, Traurigkeit und Zorn. Daraus lässt sich eine insgesamt negative Sentiment-Ausrichtung von LISETTE auf CHRYSANDER konstatieren. Über die Tooltips, die beim hovern mit der Maus angezeigt werden, kann man jedoch relativierend feststellen, dass es sich nur um wenige absolute Wörter handelt. Beide Figuren unterhalten sich also nur geringfügig im Drama, dann jedoch eher basierend auf den genannten Emotionen.

10 Fallbeispiele

Im nun letzten Kapitel werden drei Fallbeispiele diskutiert, die aufzeigen wie die SA für literaturwissenschaftliche Interpretation genutzt werden kann. Analog zur bisherigen Forschung werden dazu bekannte inhaltliche Befunde der Dramen genutzt und überprüft, inwiefern sich diese konsistent zu den Ergebnissen der SA verhalten. Dieses Vorgehen ist als kritisch zu betrachten, da man möglicherweise dazu neigt, explizit Zusammenhänge zu betrachten, die die eigene SA bestätigen. Ferner wird davon ausgegangen, dass die SA nur dann korrekt funktioniert, wenn triviale literaturwissen-

schaftliche bestätigt werden. Die grundsätzliche Idee ist für die anfängliche Analyse des Einsatzes von SA in der Dramenanalyse zunächst legitim; man sollte jedoch beachten, dass auch abweichende Aussagen einen Erkenntnisgewinn für die Literaturwissenschaft haben können und möglicherweise die Interpretation anregen. Die vorliegenden Fallbeispiele stellen insgesamt keine systematische Evaluation der SA oder eine ausgearbeitete literaturwissenschaftliche Interpretation dar. Es sollen lediglich informell an Beispielen erste Nutzungsmöglichkeiten exploriert werden. Zukünftige Projekte, die explizite professionelle Mitarbeit von Literaturwissenschaftlern nutzt, können den Einsatz der SA mit den produzierten Tools noch einmal konkreter im Kontext der literaturwissenschaftlichen Forschung untersuchen.

10.1 Fallbeispiel 1: Polaritäten im Aktverlauf

Für die Dramen von Lessing gilt zumeist, dass diese mit eine üblichen Einführung im ersten Akt beginnen und sich die Handlung dann von Akt zu Akt immer weiter über Intrigen und Ähnliches zuspitzt bis in den 5. Akt, die dann bei Tragödien mit einer Katastrophe enden (z.B. den Selbstmord Emilias in Emilia Galotti) oder in Komödien mit einer plötzlichen positiven Auflösung. Auf Basis dieser Informationen kann man die Annahme formulieren, dass der Polaritätsverlauf von Akt zu Akt negativer ausgeprägt sein sollte, insbesondere bei Tragödien.

In der Tat bestätigen einige Analysen der SA diese Annahme. Es wird die gewichtete Polarität normalisiert an der Anzahl der Wörter des jeweiligen Aktes im Aktverlauf für alle Dramen mit mehr als einem Akt betrachtet. Anbei als Beispiel der Aktverlauf zur Komödie „Der Freigeist“ und zur Tragödie „Emilia Galotti“

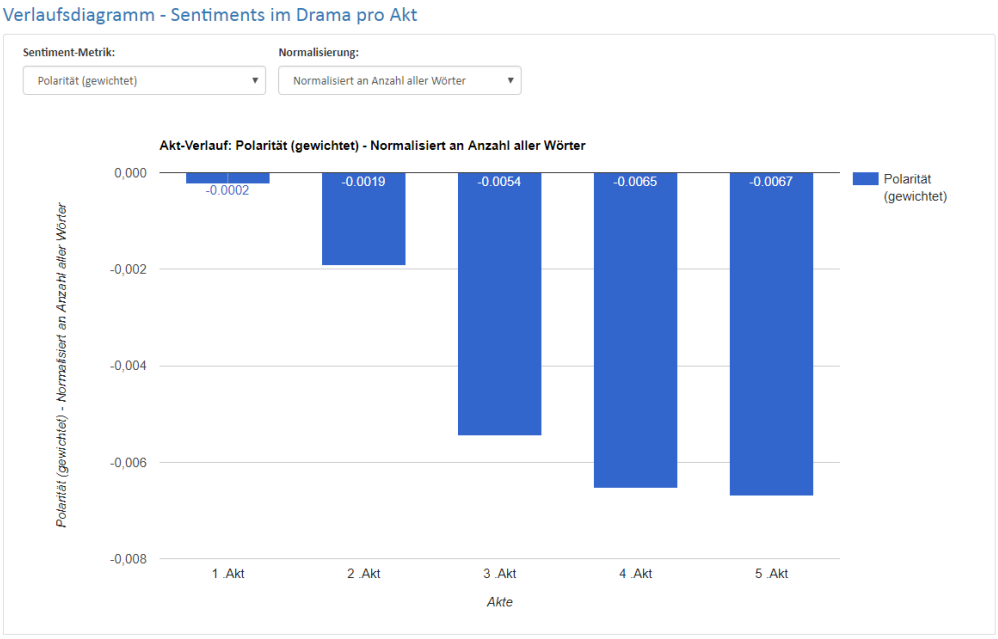


Abbildung 60: Polaritäten im Akt-Verlauf bei „Der Freigeist“

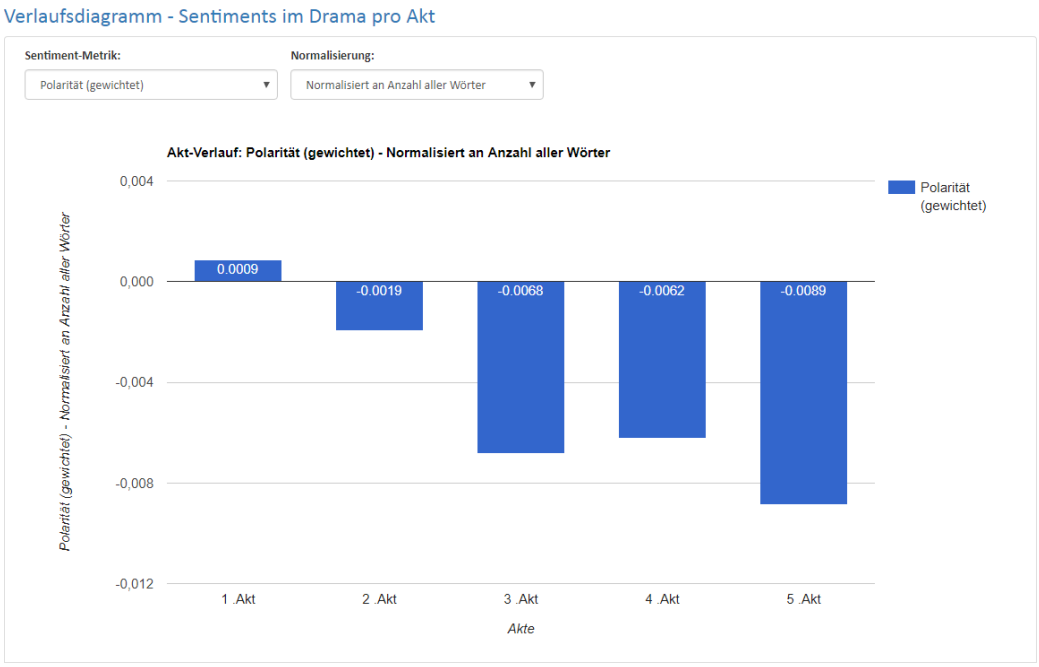


Abbildung 61: Polaritäten im Aktverlauf bei „Emilia Galotti“

Man erkennt für beide Dramen, dass die Annahme des immer negativer werdenden Aktverlaufs mit den SA-Ergebnissen pro Akt übereinstimmt. Dieser grobe Verlauf kann für alle Dramen außer Der Misogyn und Miss Sara Sampson bestätigt werden. Auf Basis dieser Erkenntnisse kann man sich nun aus literaturwissenschaftlicher Sicht mit der Frage beschäftigen, welche inhaltlichen Besonderheiten dieser Dramen zu diesem Ergebnis führen.

10.2 Fallbeispiel 2: Marinelli in Emilia Galotti

Die Figur Marinelli im Drama Emilia Galotti gilt als der klassische „Bösewicht“ und Intrigant des Dramas (Pelster, 2017). Auf Basis dieser Feststellung wird nun über das Front-End mit den berechneten SA-Metriken der gewählten Methodik untersucht, inwiefern die SA dieses Bild bestätigt.

In der Tat können vereinzelte Indikatoren gefunden werden, die Marinelli als negativ konnotierte Figur im Drama bestätigen. Zunächst kann man in der Sprecheranalyse für das ganze Drama für die Polarität (gewichtet) einen deutlichen Ausschlag in das Negative erkennen.

Abbildung 62: Polarität für die Figur Marinelli in Emilia Galotti

Im direkten Vergleich aller Figuren kann man erkennen, dass Marinelli absolut betrachtet insgesamt den am negativsten konnotierten Redeanteil aller Figuren hat.

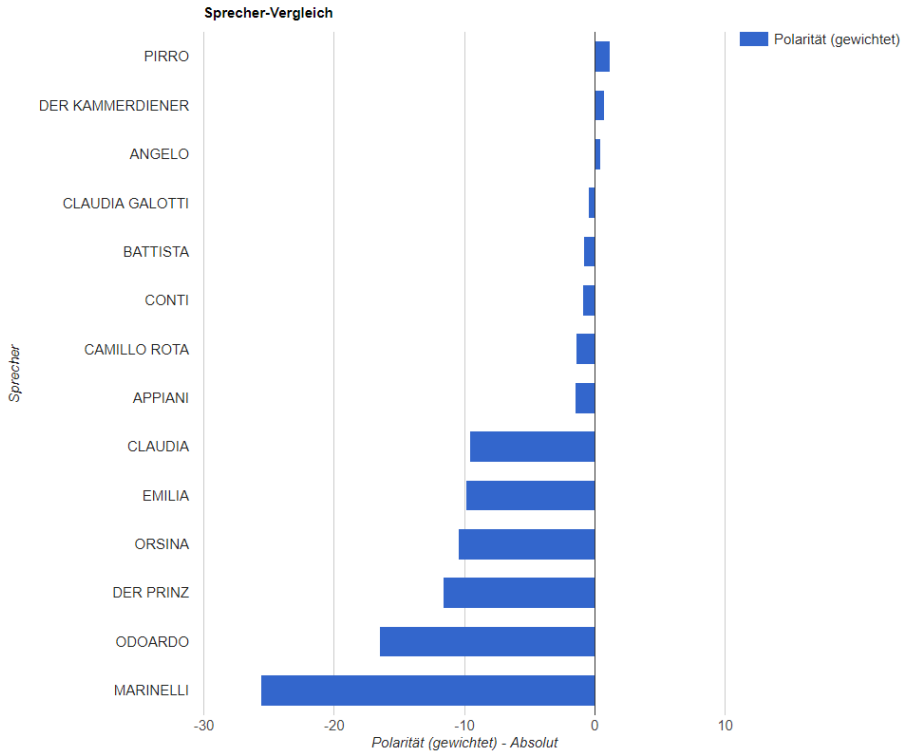


Abbildung 63: Sprecher-Vergleich für Polaritäten in Emilia Galotti

Dieses Ergebnis relativiert sich jedoch wieder, wenn man sich die gleiche Grafik normalisiert an der Anzahl der Wörter ausgeben lässt. Dann befindet sich Marinelli insgesamt lediglich im Durchschnitt. Das heißt absolut betrachtet ist Marinelli in seiner Rede zwar stark negativ konnotiert, diese starke Konnotation kommt aber vor allem durch den höheren Redeanteil generell zu Stande. Normalisiert an der Länge ist Marinelli nicht außerordentlich negativ in seiner Rede im Vergleich zu anderen Figuren. Bei den Sprecher-Beziehungen nun kann man in der Tat erkennen, dass die meisten Hauptfiguren eine starke negative Beziehung zu Marinelli aufweisen, meist sowohl absolut als auch normalisiert und häufig ist diese Beziehung die negativste. Als Beispiel wird hier die Sentiment-Beziehung des Ausgangssprechers CLAUDIA (Claudia Galotti) auf den Zielsprecher MARINELLI über die Metrik Polarität (gewichtet) und normalisiert an der Länge gezeigt:

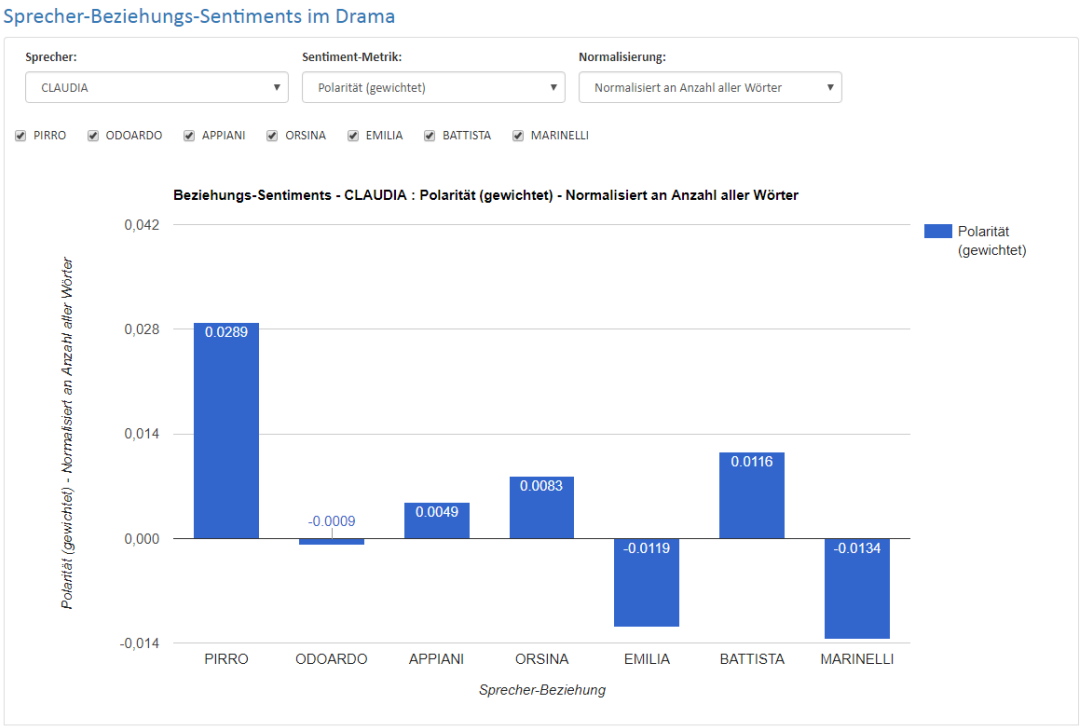


Abbildung 64: Sprecher-Beziehungspolaritäten aus Sicht der Figur Claudia

Betrachtet man Marinelli als Ausgangssprecher so kann man ebenso einige nachvollziehbare Sentiment-Beziehungen feststellen.

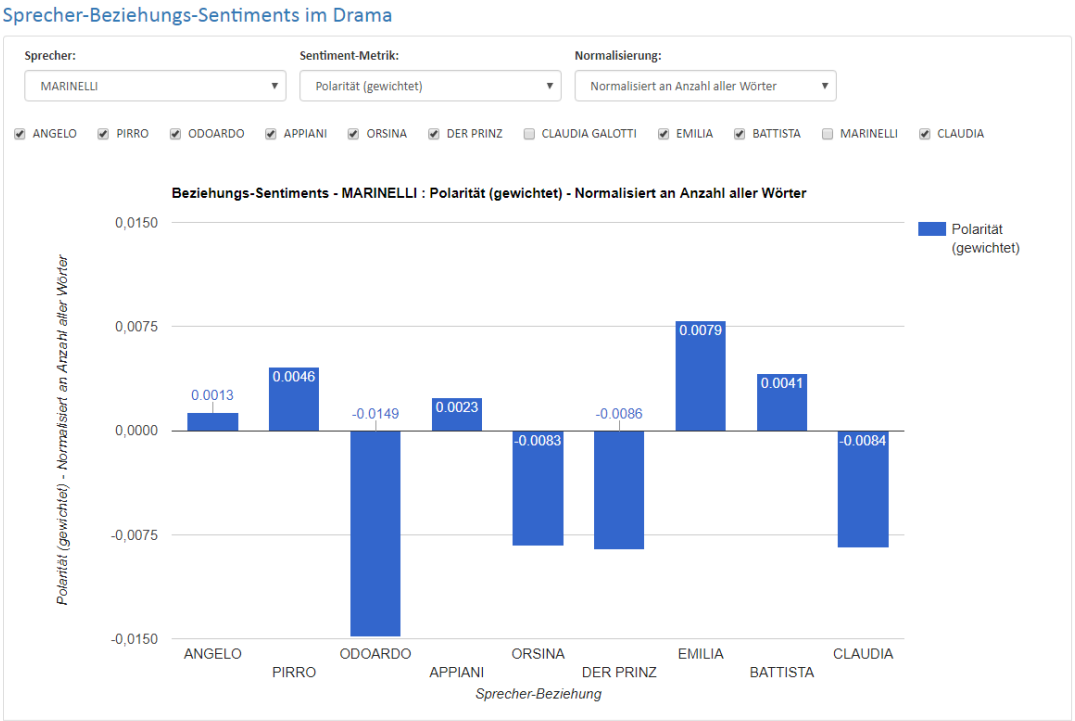


Abbildung 65: Sprecherbeziehungspolaritäten aus Sicht der Figur Marinelli

Zu den meisten Hauptfiguren bestehen aus der Sicht Marinellis negative Sentiment-Beziehungen. Die Ausnahmen hierzu passen weitestgehend zur Handlung des Dramas. Sowohl Angelo als auch Pirro und Battista sind Komplizen der intriganten und böswilligen Pläne Marinellis, weswegen die positive Sentiment-Beziehung als Nachweis für die Komplizenschaft auf Sentiment-Ebene betrachtet werden kann.

Auch die positive Beziehung zu Emilia ist erwartungskonform. Marinelli und Emilia treten nur in einer Szene in einen Dialog, welcher gemäß der gewählten Heuristik notwendig zur Kalkulation von Sentiment-Beziehungen ist. Die Szene ist die 4. Szene des 3. Aktes. Hier treten Emilia und Marinelli aufeinander und Marinelli spielt ihr vor sich für sie einsetzen zu wollen (Pelster, 2017, S. 16), verhält sich also übertrieben freundlich und höflich was von der SA gemäß der gewählten Daten bestätigt wird.

Mit den genannten zwei Fallbeispielen sollen Möglichkeiten des anfänglichen Einsatzes des SA-Tools aufgezeigt werden. Viele Funktionen wie die Emotions-Metriken wurden in den Fallbeispielen nicht angesprochen. Durch weitere Projekte in Zusammenarbeit mit der Literaturwissenschaft können weitere komplexe Anwendungsszenarien exploriert werden.

11 Diskussion und Ausblick

In der nun abschließenden Diskussion wird das Gesamtprojekt zusammengefasst und vor allem Grenzen und Anknüpfungspunkte besprochen. Eine detaillierte Diskussion zu den Arbeitspaketen der Annotationsstudie und der SA-Evaluation findet man in Kapitel 7.3 und 8.4, weswegen diese an dieser Stelle nicht wiederholt werden.

Es wurde ein mehrteiliges Projekt präsentiert, das den Einsatz von SA in einem Beispielkorpus von deutschsprachigen Dramen von Lessing untersucht. Nach Analyse der Literatur und der zur Verfügung stehenden Ressourcen wurden verschiedene Lexikon-basierte Verfahren implementiert. Es wurden Optionen wie das verwendete Lexikon, Lemmatisierung, Lexikonerweiterungen und -kombinationen umgesetzt. In einem Back-End werden verschiedene SA-Metriken für zahlreiche Dramen-Ebenen kalkuliert. Über eine erste informelle Vokabular-basierte Evaluation konnten erste Wortschatz-bezogene Erkenntnisse gesammelt werden. Es wurde ein annotiertes Test-Korpus aus 200 Repliken erstellt und die Leistung von Verfahren anhand dieses Kor-

pus evaluiert. Insbesondere die Lexikonerweiterung mit historischen und linguistischen Varianten und das Lexikon SentiWS konnten hohe Genauigkeitsraten erzielen. In einem webbasierten Front-End wird die Exploration von Polaritäts- und Emotions-Metriken auf verschiedenen Ebenen eines Dramas ermöglicht. Insgesamt konnten die bisherigen Arbeiten zur SA in der Dramenanalyse von Mohammad (2011) und Nalisnick und Baird (2013) erfolgreich weiter geführt werden. Das Projekt stellt durch die erstmalige Erstellung eines annotierten Test-Korpus, die statistische Analyse von Annotationsverhalten, die systematische Evaluation verschiedener Lexikon-basierter Verfahren und der Entwicklung einer umfangreichen interaktiven Visualisierungs-Anwendung einen deutlichen Mehrwert für die bisherige Forschung dar, auf dem verwandte Projekte aufbauen können.

Man kann konstatieren, dass die Teilergebnisse der Einzel-Studien und -Pakete aufeinander aufbauen. Die erhöhte Problematik in der Annotation der Repliken setzt sich in einer, im Vergleich zu anderen SA-Gebieten, schwächeren Genauigkeit bei der Prädiktion fort. Die übermäßige Wahrnehmung von negativen Repliken bei der Annotation, setzt sich in einer übermäßigen Prädiktion negativer Repliken der verwendeten Algorithmen fort und muss auch stets bei der Interpretation beachtet werden. Es ist Aufgabe der Literaturwissenschaft, sich mit dem erhöhten Auftreten von Negativität in dem gewählten Korpus auseinanderzusetzen.

Obschon systematisch zahlreiche Lexikon-basierte Verfahren untersucht wurden, bleiben einige generelle Probleme des Lexikon-Einsatzes bestehen und die SA bietet zahlreiche weitere Verfahren um diese möglicherweise zu umgehen. Die Kritik für Lexikon-basierte Methoden wurde bereits in Kapitel 2.2 angesprochen und trifft auch für den vorliegenden Anwendungsfall zu. Auch im Bereich der Lexikon-basierten Verfahren wurden im vorliegenden Projekt noch nicht alle Möglichkeiten auf ihre Nutzbarkeit untersucht. Es wurde beispielsweise nicht der Einfluss von Negationen und Verstärker-Wörtern oder Valenz-Shiftern untersucht. Außerdem wurde eine primitive Lexikon-Kombination im Vergleich zu Emerson und Declerck (2014) gewählt. Diese könnte durch komplexe mathematische Verfahren, die zum Beispiel auch die Länge eines Lexikons beachten, verbessert werden. Als Lexikonerweiterungs-Option steht noch die Möglichkeit der Erweiterung von Synonymen zur Verfügung. Auch die semi-automatische Übersetzung von bislang noch nicht übersetzten englischsprachigen Le-

xika kann ein Ansatz zur Optimierung sein. Die in der Vokabular-basierten Evaluation erstellten Wortschatz-Sammlungen des Korpus können auch zur Verbesserung der Identifikation von SBWs genutzt werden. Kenntnisreiche Sprachwissenschaftler können beispielsweise erkannte oder nicht erkannte Wörter händisch bezüglich Polarität beurteilen und ausbessern um eine präzisere Prädiktion zu erlangen.

Neben Negationen und Verstärker-Wörtern sind auch andere Regel-basierte Methoden möglich, die explizit an das Konzept des Dramas angepasst sind. Literaturwissenschaftliche Expertise könnte einen Beitrag dazu leisten, welche speziellen textuellen oder strukturellen Elemente explizit beachtet werden können um die Zuordnung der Polarität mittels Regeln zu steuern. Ein Beispiel dazu sind Regie-Anweisungen. Häufig sind in diesen exakte Sentiment-Informationen zum Verhalten der Figur in der Form *SPRECHER: (zornig) Sprechakt*. Bislang werden solche Regieanweisungen in der vorliegenden Arbeit als herkömmliche Wörter der Replik betrachtet. Tatsächlich handelt es sich aber im Prinzip um eine explizite Annotation des Autors zum Sentiment dieser Replik und man könnte regelbasiert unabhängig vom sonstigen Inhalt annehmen, dass diese negativ ist oder sie negativer gewichten. Weitere interdisziplinäre Projekte können derartige Besonderheiten und den Nutzen für die korrekte SA weiter untersuchen.

Aufgrund der veralteten und poetischen Sprache der Textgrundlage ist die Tauglichkeit der Lexika begrenzt. Mit fortgeschrittenen Methoden, die maschinelles Lernen in den Ansatz integrieren, können möglicherweise bessere Ergebnisse erzielt werden. Ansätze mit maschinellem Lernen wurden in der vorliegenden Studie bislang ausgeschlossen, da für die meisten Methoden annotierte Test-Korpora benötigt werden, die zu Beginn nicht vorlagen. Die hier erlangten annotierten Repliken können als Anfang für eine größere Annotationsstudie betrachtet werden, um ausreichende Auszeichnungen auf Satz- oder Repliken-Ebene zu erlangen und somit simple ML-Algorithmen wie Naive Bayes oder SVM zu trainieren. Das momentane Korpus ist dafür eher noch zu klein und die Annotationen, vor allem für positive Repliken sind zu wenig übereinstimmend. Wie das Akquirieren von Annotationen verbessert werden kann, wurde bereits in Kapitel 7.3 angesprochen.

Um die problematische veraltete Sprache besser zu kontrollieren erscheint es als vielversprechend mittels bekannter SA-Methoden ein domänenspezifisches Lexikon zu erstellen (siehe Kapitel 2.1 und 2.2). Dadurch können beispielsweise unbekannte Wör-

ter identifiziert werden, die oft zusammen mit bekannten positiven oder negativen Repliken vorliegen und das SA-Lexikon sprachspezifisch erweitert werden. Generell sollten zukünftige Arbeiten vorrangig die Möglichkeit hybrider Ansätze explorieren (siehe Kapitel 2.1). Hierzu ist nicht zwingend ein annotiertes Trainings-Korpus notwendig. Die bislang entwickelten Lexikon-Verfahren können beispielsweise eingesetzt werden, um eine ausreichend große Menge an übereinstimmend negativen und positiven Repliken oder Sätzen zu akquirieren. Diese können manuell überprüft werden und dann zum Training eines ML-Algorithmus genutzt. Da sich ML-Methoden meist gegenüber reinen Lexikon-Methoden durchsetzen, wird die Integration von ML in die SA auf der Dramenanalyse als primärer nächster Arbeitsschritt betrachtet.

Als weiterer nächster Anknüpfungspunkt wird die Analyse von weiteren Dramen aus anderen Epochen und von anderen Dramatikern empfohlen. Das Back-End und das Tool wurden derart entwickelt, dass sämtliche von Text Grid akquirierte korrekt annotierte Dramen verarbeitbar und nutzbar sind. Dieser Umstand kann aufgegriffen werden, um direkt weitere Dramen und Autoren zu analysieren. Möglicherweise lassen sich Dramen finden, die in Zusammenhang mit der gelieferten SA weitere nützliche Ergebnisse produzieren, die direkt erfolgreich interpretiert werden können. Des Weiteren ist es denkbar, dass andere Epochen und Autoren näher an der Sprache der verwendeten SA-Lexika liegen. Über weiterführende Projekte können dann die Funktionen von Dramen-, Autoren- oder Epochenvergleichen eingebaut werden, die bislang nicht im Tool enthalten sind. So kann die SA im übergeordneten globalen Kontext untersucht werden.

Die ersten Visualisierungsansätze von Mohammad (2011) und Nalisnick und Baird (2013) werden in der vorliegenden Arbeit aufgegriffen, aber deutlich ausgebaut. Die SA-Erweiterung von Katharsis ist nach Kenntnisstand des Autors das erste frei verfügbare interaktive Tool, das die Möglichkeit der visuellen Exploration von Dramen anbietet. Die Visualisierungen wurden nach eigenem Ermessen und gemäß der zur Verfügung stehenden technischen Optionen implementiert. Tatsächlich wurde es aber ohne eine größere Anforderungserhebung oder Konzeption implementiert. Generell ist der Fokus momentaner SA-Forschung weiterhin die reine Prädiktion von Polarität, die Analyse von Visualisierungsmöglichkeiten ist noch vergleichsweise selten. Die Integration von Nutzer-zentrierten Methoden der Usability- und User Experience-Forschung

können einen wesentlichen Beitrag zur Verbesserung der Nutzbarkeit und des Nutzer-Erlebnisses leisten. Das momentane Tool wird eher als erster Ansatz betrachtet, um weitere optimierte, verständliche und visuell ansprechende Darstellungen der SA-Daten zu entwickeln und zu explorieren. Speziell für den Bereich des Webs wird empfohlen, fortgeschrittene Technologien der Datenvisualisierungen jenseits der genutzten Google Charts-Library wie D3 auf ihren gewinnbringenden Einsatz zu untersuchen.

Grundsätzlich ist anzumerken, dass das Tool als Experten-Tool für Literaturwissenschaftler gedacht ist. Aus diesem Grund ist es empfehlenswert bei zukünftigen nutzer-zentrierten Anschlussstudien diese Personengruppe explizit miteinbeziehen. Das entwickelte Tool kann über Usability-Tests und ähnliches genutzt werden um Anforderungen und Wünsche dieser speziellen Zielgruppe zu erheben und umzusetzen.

Eine zentrale Erkenntnis des gesamten Projekts ist die übermäßige Häufigkeit negativer Repliken sowohl aus Annotatorensicht als auch über die meisten verwendeten Repliken. Auch die verwendete gewichtete Metrik weist diese Überbewertung von Repliken als negativ auf. Obschon die Annotation belegt, dass Negativität in dem Korpus vorherrschend ist, macht dies die Interpretation von Daten problematisch, da beispielsweise Akte oder Sprecher grundsätzlich negativ bewertet sind. Dieser Umstand ist stets bei der Interpretation insofern zu beachten, dass direkte Aussagen über einzelne Ebenen in der Form: Akt X ist negativ, Drama X ist positiv etc. meist nicht korrekt sind. Es wird empfohlen stets den Vergleich heranzuziehen, also zum Beispiel zwischen Akten zu betrachten, welcher weniger negativ ist und so auch bezüglich anderer Ebenen und Analysen zu verfahren. Aufgrund des Übergewichts von negativen Repliken kann ein vergleichsweise weniger negativer Akt als positiv betrachtet werden. In Kapitel 10 wurde gezeigt, wie trotz dieser Problematik sinnvolle Interpretationen durchgeführt werden können. Es wurden spezielle Normalisierungsmetriken implementiert, um die Vergleichbarkeit von Einheiten und Ebenen zu ermöglichen und zu vereinfachen. Zukünftige Projekte können dieses Thema jedoch explizit adressieren indem beispielsweise über statistische Analysen untersucht wird, ob statt der konkreten SA-Kalkulation Mittelwerte oder ähnliches definiert werden können zur heuristischen Bestimmung von Polaritäten gemäß dieser. Auch liegt diese Überbewertung nicht für jede Metrik vor, jedoch erzielen derartige Metriken bessere Evaluationsergebnisse. Es ist mit der vorliegenden Arbeit möglich auch Metriken auszutesten, die eine

ausgeglichenere Prädiktion von negativen und positiven Repliken vornehmen und zu untersuchen, ob diese für die Gesamtinterpretation gewinnbringender sind.

Bezüglich der Metriken-Kalkulation sind vereinzelte Aspekte kritisch zu betrachten. Die Berechnung von Sprecherbeziehungen folgt einer sehr simplen Heuristik nach Nalisnick und Baird (2013). Diese hat auch den Nachteil, dass weniger tatsächlich aufeinander bezogene Relationen berechnet werden, sondern die Sichtweise einer Person auf die andere. Es wird eine Entfernung vom Fokus auf den reinen Dialog zwischen Personen empfohlen. So können Erwähnungen der Namen von Personen genutzt werden um Beziehungsmetriken unabhängig davon zu kalkulieren ob Personen aufeinander treffen. Dies ist zielführend, da Beziehungen auch dann bestehen wenn Personen nicht aufeinander treffen, sondern beispielsweise nur übereinander reden. Die korrekte Kalkulation von Beziehungen wird als sehr anspruchsvolles Problem erachtet, das in separaten Studien untersucht werden sollte.

Es wurde bereits in Kapitel 3 bezüglich der Forschungsfrage angemerkt, dass diese lediglich zur Orientierung dient und über die vorliegende Arbeit nicht vollständig beantwortet werden kann. Erste Fallbeispiele aus Kapitel 10 legen jedoch nahe, dass die SA in der Tat gewinnbringend zur literaturwissenschaftlichen Interpretation eingesetzt werden kann. Um sich jedoch weiterführend mit der Forschungsfrage auseinanderzusetzen, ist es notwendig, die Zusammenarbeit mit Literaturwissenschaftlern zu suchen. Auf diese Weise können qualifizierte Analysen vorgenommen werden und der Einsatz der SA präziser untersucht werden als es im vorliegenden Projekt möglich war.

Abschließend sei noch, auch bezogen auf den zuletzt angesprochenen Punkt erwähnt, dass das grundsätzliche Schema in der Forschung, zu testen, ob die SA bekannte literarische Zusammenhänge bestätigt, um den Nutzen von SA zu bekräftigen, sehr zweifelhaft ist. Zwar ist diese Art der informellen Evaluation zum jetzigen Stand der Forschung berechtigt. Es muss jedoch beachtet werden, dass auch die Möglichkeit besteht, dass die SA auf Zusammenhänge hinweist, die für die literarische Interpretation bislang noch nicht bekannt waren. Es sollte angestrebt werden, dass die SA für literarische Texte nicht nur zur Bestätigung genutzt wird sondern auch bisherige Annahmen erweitert oder zu ganz neuen Erkenntnissen führt.

Mit dem präsentierten Projekt konnte ein erster großer Beitrag zur Zusammenführung von SA und quantitativer Dramenanalyse geleistet werden. Weitere interdisziplinäre Forschung und Projekte zwischen der Medieninformatik und der Literaturwissenschaft sind jedoch nötig um die Möglichkeiten und den Nutzen der SA für die literaturwissenschaftliche Interpretation weiter zu explorieren. Mit der vorliegenden Arbeit werden erste Arbeitspakete und Resultate geliefert, um diese Zusammenarbeit anzustoßen und auszubauen.

Literaturverzeichnis

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

Agarwal, B., Mittal, N., Bansal, P., & Garg, S. (2015). Sentiment analysis using common-sense and context information. *Computational intelligence and neuroscience*, 2015, 30.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media* (pp. 30-38). Association for Computational Linguistics.

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.

Alm, C. O. & Sproat, R. (2005a). Emotional sequencing and development in fairy tales. In *International Conference on Affective Computing and Intelligent Interaction* (pp. 668-674). Springer Berlin Heidelberg.

Alm, C. O., & Sproat, R. (2005b). Perceptions of emotions in expressive storytelling. In *Ninth European Conference on Speech Communication and Technology*.

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 579-586). Association for Computational Linguistics.

Antoine, J. Y., Villaneau, J., & Lefeuvre, A. (2014). Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In *EACL 2014* (pp. 10-p).

Asghar, M. Z., Khan, A., Ahmad, S., & Kundi, F. M. (2014). A review of feature extraction in sentiment analysis. *Journal of Basic and Applied Scientific Research*, 4(3), 181-186.

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.

Alt, Peter-André. (1994). *Tragödie der Aufklärung. Eine Einführung*. Tübingen: UTB Verlag.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database* [CD-ROM]. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

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. (2009). Rethinking Sentiment Analysis in the News: from Theory to Practice and back. *Proceeding of WOMSA*, 9.

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

Basile, V., & Nissim, M. (2013, June). Sentiment analysis on Italian tweets. In *WASSA@ NAACL-HLT* (pp. 100-107).

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.

Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A., & Reynar, J. (2008). Building a sentiment summarizer for local service reviews. *In WWW workshop on NLP in the information explosion era* (Vol. 14, pp. 339-348).

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

Bosco, C., Allisio, L., Mussa, V., Patti, V., Ruffo, G., Sanguinetti, M., & Sulis, E. (2014, May). Detecting happiness in Italian tweets: Towards an evaluation dataset for sentiment analysis in Felicitta. In *Proc. of the 5th International Workshop on Emotion, Social Signals, Sentiment and Linked Opena Data, ESSSLOD* (pp. 56-63).

Botting, F. (1996). *Gothic (The New Critical Idiom)*. New York, USA: Routledge.

Burghardt, M., Dennerlein, K., Schmidt, T., Mühlenfeld, J. & Wolff, C. (2016). Katharsis – Ein Werkzeug für die quantitative Dramenanalyse. In *CLARIN-D Forum CA3*. Retrieved from https://www.clarin-d.de/de/konferenz-abstracts/369-katharsis-ein-werkzeug-fuer-die-quantitative-dramenanalyse

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*.

Carvalho, P., Sarmento, L., Silva, M. J., & De Oliveira, E. (2009, November). Clues for detecting irony in user-generated contents: oh...!! it's so easy;-. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion* (pp. 53-56). ACM.

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

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).

Clematide, S., Gindl, S., Klenner, M., Petrakis, S., Remus, R., Ruppenhofer, J., ... & Wiegand, M. (2012). MLSA-A Multi-layered Reference Corpus for German Sentiment Analysis. In *LREC* (pp. 3551-3556).

Clematide, S. & Klenner, M. (2010). Evaluation and extension of a polarity lexicon for German. In *Proceedings of the First Workshop on Computational Approaches to Subjectivity and Sentiment Analysis* (pp. 7-13).

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.

De Smedt, T. & Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning Research*, 13: 2031–2035.

De Fortuny, E. J., De Smedt, T., Martens, D., & Daelemans, W. (2012). Media coverage in times of political crisis: A text mining approach. *Expert Systems with Applications*, 39(14), 11616-11622.

Denecke, K. (2008). Using sentiwordnet for multilingual sentiment analysis. *In Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on* (pp. 507-512). IEEE.

Ding, X., Liu, B., & Yu, P. S. (2008). 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.

Ding, X., Liu, B., & Zhang, L. (2009, June). Entity discovery and assignment for opinion mining applications. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1125-1134). ACM.

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

Eckle-Kohler, J., Kluge, R., & Gurevych, I. (2015). On the Role of Discourse Markers for Discriminating Claims and Premises in Argumentative Discourse. In *EMNLP* (pp. 2236-2242).

Eger, S., Gleim, R., & Mehler, A. (2016). Lemmatization and Morphological Tagging in German and Latin: A Comparison and a Survey of the State-of-the-art. In *LREC*.

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.

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).

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.

Feinstein, A. R., and Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43, 543-549.

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

Fellbaum, C. (1998) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fick, M. (2016). *Lessing-Handbuch. Leben – Werk – Wirkung*. Stuttgart; Weimar: Verlag J.B. Metzler.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5), 378-382.

Frederking, V. (2016). *Literarische Texte. Literarische und nicht-literarische Texte*. Retrieved from http://www.br.de/alphalernen/faecher/deutsch/2-literarische-nicht-literarische-texte-literatur-102.html

Fucks, W. & Lauter, J. (1965). Mathematische Analyse des literarischen Stils. In Kreuzer, H. & Gunzenhäuser, F. (Hrsg.), *Mathematik und Dichtung*, (S. 107-122). München: Nymphenburger Verlagshandlung.

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.

Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013). Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks* (pp. 27-38). ACM.

Gotthold Ephraim Lessing. (o. J.). In *Wikipedia*. Retrieved from https://de.wikipedia.org/wiki/Gotthold_Ephraim_Lessing

Gwet, K. L. (2011). *On The Krippendorff's Alpha Coefficient*. Retrieved from http://www.agreestat.com/research_papers/onkrippendorffalpha.pdf

Hamp, B., & Feldweg, H. (1997). Germanet-a lexical-semantic net for german. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications* (pp. 9-15).

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics* (pp. 174-181). Association for Computational Linguistics.

Hogenboom, A., Frasincar, F., De Jong, F., & Kaymak, U. (2015). Using rhetorical structure in sentiment analysis. *Communications of the ACM*, 58(7), 69-77.

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

Hölzer, M., Scheytt, N., & Kächele, H. (1992). Das „Affektive Diktionär Ulm "als eine Methode der quantitativen Vokabularbestimmung. In *Textanalyse* (pp. 131-154). VS Verlag für Sozialwissenschaften.

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.

Hu, X., Tang, J., Gao, H., & Liu, H. (2013). Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 607-618). 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.

Ilsemann, H. (2005). Some statistical observations on speech lengths in Shakespeare's plays. *Shakespeare Jahrbuch*, 141, 158–68.

Ilsemann, H. (2008). More statistical observations on speech lengths in Shakespeare's plays. *Literary and Linguistic Computing*, 23(4), 397-407.

Jannidis, F., Reger, I., Zehe, A., Becker, M., Hettinger, L. & Hotho, A. (2016). *Analyzing Features for the Detection of Happy Endings in German Novels*. arXiv preprint arXiv:1611.09028.

Joyce, M. (2013). *Picking the best Intercoder reliability statistic for your digital activism content analysis.* Retrieved from http://digital-activism.org/2013/05/picking-the-best-intercoder-reliability-statistic-for-your-digital-activism-content-analysis/

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

Jurish, B. (2012). *Finite-state Canonicalization Techniques for Historical German*. PhD thesis, Universität Potsdam, 2012 (defended 2011). URN urn:nbn:de:kobv:517-opus-55789.

Kaji, N., & Kitsuregawa, M. (2007). Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. In *EMNLP-CoNLL* (pp. 1075-1083).

Kakkonen, T. & Kakkonen, G. G. (2011). SentiProfiler: creating comparable visual profiles of sentimental content in texts. In *Proceedings of Language Technologies for Digital Humanities and Cultural Heritage* (pp. 62-69).

Kamps, J., Marx, M., Mokken, R. J., & De Rijke, M. (2004). Using WordNet to Measure Semantic Orientations of Adjectives. In *LREC* (Vol. 4, pp. 1115-1118).

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.

Kessler, J. S., Eckert, M., Clark, L., & Nicolov, N. (2010). The ICWSM 2010 JDPA sentiment corpus for the automotive domain. In *4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC),* Washington, DC.

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 Linguistic*s (p. 1367). Association for Computational Linguistics.

Klinger, R., Suliya, S. S., & Reiter, N. (2016). Automatic Emotion Detection for antitative Literary Studies. In *Digital Humanities 2016: Conference Abstracts* (pp. 826-828).

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.

Krippendorff, K. (2011). *Computing Krippendorff's Alpha-Reliability.* Retrieved from http://repository.upenn.edu/asc_papers/43

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.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

Lang, P. J. (1980). Behavioral treatment and bio-behavioral assessment: Computer applications. In J. B. Sidowski, J. H. Johnson, & T. A. Williams (Eds.), *Technology in mental health and delivery systems* (pp. 119-137). Norwood, NJ: Ablex.

Lemma (Lexikographie). (o. J.). In *Wikipedia.* Retrieved from https://de.wikipedia.org/wiki/Lemma_(Lexikographie)

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.

Malo, P., Sinha, A., Takala, P., Ahlgren, O., and Lappalainen, I. (2013). Learning the Roles of Directional Expressions and Domain Concepts in Financial News Analysis. In *Proceedings of IEEE International Conference of Data Mining workshops (ICDM SENTIRE).* IEEE Press.

Marchetti, A., Sprugnoli, R. & Tonelli, S. (2014). Sentiment Analysis for the Humanities: The Case of Historical Texts. In *DH 2014*. Retrieved from http://dharchive.org/paper/DH2014/Paper-220.xml

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

Maynard, D., & Bontcheva, K. (2016). Challenges of Evaluating Sentiment Analysis Tools on Social Media. In *LREC*.

Maynard, D., & Greenwood, M. A. (2014). Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis. In *LREC* (pp. 4238-4243).

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.

Mellmann, K. (2007). *Emotionalisierung – Von der Nebenstundenpoesie zum Buch als Freund. Vol. 4. Poetogenesis - Studien zur empirischen Anthropologie der Literatur.* Münster: Mentis.

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. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. *In Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 26-34). Association for Computational Linguistics.

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

Momtazi, S. (2012). Fine-grained German Sentiment Analysis on Social Media. In *LREC* (pp. 1215-1220).

Monz, C., & De Rijke, M. (2001). Shallow morphological analysis in monolingual information retrieval for Dutch, German, and Italian. In *Workshop of the Cross-Language Evaluation Forum for European Languages* (pp. 262-277). Springer, Berlin, Heidelberg.

Mozetič, I., Grčar, M., & Smailović, J. (2016). Multilingual Twitter sentiment classification: The role of human annotators. *PloS one*, 11(5), e0155036.

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).

Nalisnick, E. T., & Baird, H. S. (2013). Character-to-character sentiment analysis in shakespeare's plays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 479–483).

Nathan der Weise. (o. J.). In *Wikipedia*. Retrieved from https://de.wikipedia.org/wiki/Nathan_der_Weise

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

Ortner, H. (2014). *Text und Emotion. Theorie, Methode und Anwendungsbeispiele emotionslinguistischer Textanalyse*. Tübingen: Narr Verlag.

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.

Palanisamy, P., Yadav, V., & Elchuri, H. (2013). Serendio: Simple and Practical lexicon based approach to Sentiment Analysis. In *proceedings of Second Joint Conference on Lexical and Computational Semantics* (pp. 543-548).

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.

Peleja, F., Santos, J., & Magalhães, J. (2014). Reputation analysis with a ranked sentiment-lexicon. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (pp. 1207-1210). ACM.

Pelster, T. (2017). *Emilia Galotti. Reclam Lektüreschlüssel XL*. Stuttgart: Philipp Reclam.

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

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Theories of emotion*, 1(3-31), 4.

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

Qiu, G., Liu, B., Bu, J., & Chen, C. (2009, July). Expanding domain sentiment lexicon through double propagation. In *IJCAI* (Vol. 9, pp. 1199-1204).

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.

Qiu, G., Liu, B., Bu, J., & Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1), 9-27.

Quasthoff, U. (2010). Deutsches Kollokationswörterbuch. Berlin: deGruyter.

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.

Refaee, E., & Rieser, V. (2014). An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis. In *LREC* (pp. 2268-2273).

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

Reyes, A., Rosso, P., & Veale, T. (2013). A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1), 239-268.

Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., & Benevenuto, F. (2016). Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1), 1-29.

Rill, S., Scheidt, J., Drescher, J., Schütz, O., Reinel, D., & Wogenstein, F. (2012). A generic approach to generate opinion lists of phrases for opinion mining applications. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining* (p. 7). ACM.

Russell, J. (1980). A circumplex model of affect. In *Journal of Personality and Social Psychology*, 39, 1161–1178.

Saif, H., Fernandez, M., He, Y., & Alani, H. (2013). Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold. In: *1st Interantional Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013)*

Saif, H., Fernandez, M., He, Y., Alani, H. (2014). On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter. In: *Proc. 9th Language Resources and Evaluation Conference (LREC)* (pp. 810-817).

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.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

Shin, H., Kim, M., Jang, H., & Cattle, A. (2012, November). Annotation Scheme for Constructing Sentiment Corpus in Korean. In *PACLIC* (pp. 181-190).

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.

Solomon, M. (1971). Ein mathematisch-linguistisches Dramenmodell. *Zeitschrift für Literaturwissenschaft und Linguistik*, 1(1), 139-152.

Solomon, M. (1973). Mathematische Poetik. Frankfurt: Athenäum

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

Stone, P., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1968). The general inquirer: A computer approach to content analysis. *Journal of Regional Science*, 8(1), 113-116.

Strapparava, C., & Valitutti, A. (2004). WordNet Affect: an Affective Extension of WordNet. In *LREC* (Vol. 4, pp. 1083-1086).

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2), 267-307.

Takala, P., Malo, P., Sinha, A., & Ahlgren, O. (2014). Gold-standard for Topic-specific Sentiment Analysis of Economic Texts. In *LREC* (Vol. 2014, pp. 2152-2157).

Takamura, H., Inui, T., & Okumura, M. (2005). Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 133-140). Association for Computational Linguistics.

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.

Thielen, C. , Schiller, A., Teufel, S. & Stöckert, C. (1999). *Guidelines für das Tagging deutscher Textkorpora mit STTS*. Technical report, University of Stuttgart and University of Tübingen.

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).

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417-424). Association for Computational Linguistics.

Van de Kauter, M., Desmet, B., & Hoste, V. (2015). The good, the bad and the implicit: a comprehensive approach to annotating explicit and implicit sentiment. *Language Resources and Evaluation*, 49(3), 685-720.

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.

Volkova, E. P., Mohler, B. J., Meurers, D., Gerdemann, D., & Bülthoff, H. H. (2010, June). Emotional perception of fairy tales: achieving agreement in emotion annotation of text. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 98-106). Association for Computational Linguistics.

Waltinger, U. (2010). Sentiment Analysis Reloaded: A Comparative Study On Sentiment Polarity Identification Combining Machine Learning And Subjectivity Features. In *Proceedings of the 6th International Conference on Web Information Systems and Technologies (WEBIST '10).*

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.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2), 165-210.

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.

Winko, S. (2003). *Kodierte Gefühle: Zu einer Poetik der Emotionen in lyrischen und poetologischen Texten um 1900*. Berlin: Erich Schmidt Verlag.

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.

Zhang, L., Gosh, R., Riddhiman, D., Dekhil, M., Hsu, M. & Liu, B. (2011). *Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis*. Retrieved from http://www.hpl.hp.com/techreports/2011/HPL-2011-89.html

Zhou, Z., Zhao, W., & Shang, L. (2014). Sentiment analysis with automatically constructed lexicon and three-way decision. In *International Conference on Rough Sets and Knowledge Technology* (pp. 777-788). Springer, Cham.

Anhang – DVD

Agreement-Daten

- Auswertungen bezüglich Übereinstimmungsanalysen der Annotationsstudie

Annotationsstudie

- Annotation-Auswertung
- Fragebogen
- Test-Korpus-Annotationen

Evaluation

- Test-Korpus-Evaluation
- Vocabulary-Evaluation

Front-End

- Webanwendung → sa_selection.html zum starten

Korpus (alte Katharsis-Version)

Lessing-Dramen (Bearbeitete Lessing-Dramen)

Lessing-Dramen-roh (unbearbeitete Lessing-Dramen)

Python

- Python-Back-End

Sentiment Analysis

- Original-Lexika
- Transformierte Lexika-Dateien
- DTA-Output
- SA-Output (JSON für verwendete SA-Methode)

Stopwords

Videos

Word-Frequencies

- Vokabular-Dateien für alle Dramen und das Gesamtkorpus

Plagiatserklärung

Ich habe die Arbeit selbständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und bisher keiner anderen Prüfungsbehörde vorgelegt. Von den zu § 26 Abs. 5 der Prüfungsordnung vorgesehenen Rechtsfolgen habe ich Kenntnis. Die vorgelegten Druckexemplare und die vorgelegte digitale Version sind identisch.

_____ _____

Ort, Datum, Unterschrift