Sentimentanalyse als Instrument literaturgeschichtlicher Rezeptionsforschung
–
Ein Pilotprojekt

Keli Du, Katja Mellmann

DARIAH-DE Working Papers Nr.32

Du, Kelly; Mellmann, Katja: Sentimentanalyse als Instrument literaturgeschichtlicher Rezeptionsforschung : ein Piloprojekt
Göttingen : GOEDOC, Dokumenten- und Publikationsserver der Georg-August-Universität, 2016
(DARIAH-DE working papers 32)

Verfügbar:
PURL: http://resolver.sub.uni-goettingen.de/purl/?dariah-2019-4
URN: http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2019-4-2

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über <http://dnb.dnb.de> abrufbar.

Erschienen in der Reihe
DARIAH-DE working papers

ISSN: 2198-4670

Herausgeber der Reihe
DARIAH-DE, Niedersächsische Staats- und Universitätsbibliothek

Mirjam Blümm, Thomas Kollatz, Stefan Schmunk und Christof Schöch

---

Abstract: Die wachsende Anzahl an Retrodigitalisaten historischer Literatur- und Kulturzeitschriften stellt eine wichtige Quelle für literaturgeschichtliche Rezeptionsforschung dar. Der folgende Aufsatz berichtet von unserem Versuch, an Hand eines Probekorpus Instrumente für die Sentimentanalyse literaturkritischen Schrifttums des späten 19. Jahrhunderts zu entwickeln. Dies umfasst unter anderem eine Domänenanpassung des Sentimentwort-Lexikons, Überlegungen zur Analyseeinheit, zur Berechnung des Sentiment-Werts, zur Eignung weiterer Features für eine automatische Klassifikation und zum Umgang mit Problemen der Textsorte Rezension.

Keywords: Literaturgeschichte, Rezensionen, Literaturkritik, Wirkung, historische Rezeptionsforschung, Korpusanalyse, Sentimentanalyse, Zeitschriften, Periodika, 19. Jahrhundert

literary history, reviews, literary criticism, effect, historical reception studies, corpus analysis, Sentiment analysis, journals, periodicals, 19th century

Sentimentanalyse als Instrument literaturgeschichtlicher Rezeptionsforschung

Ein Pilotprojekt

Keli Du Katja Mellmann

Seminar für deutsche Philologie, Universität Göttingen

DARIAH-DE Working Papers

Keli Du, Katja Mellmann: „Sentimentanalyse als Instrument literaturgeschichtlicher Rezeptionsforschung". *DARIAH-DE Working Papers* Nr. 32. Göttingen: DARIAH-DE, 2019.
URN: urn:nbn:de:gbv:7-dariah-2019-4-2.

Zusammenfassung

Die wachsende Anzahl an Retrodigitalisaten historischer Literatur- und Kulturzeitschriften stellt eine wichtige Quelle für literaturgeschichtliche Rezeptionsforschung dar. Der folgende Aufsatz berichtet von unserem Versuch, an Hand eines Probekorpus Instrumente für die Sentimentanalyse literaturkritischen Schrifttums des späten 19. Jahrhunderts zu entwickeln. Dies umfasst unter anderem eine Domänenanpassung des Sentimentwort-Lexikons, Überlegungen zur Analyseeinheit, zur Berechnung des Sentiment-Werts, zur Eignung weiterer Features für eine automatische Klassifikation und zum Umgang mit Problemen der Textsorte Rezension.

Schlagwörter

Literaturgeschichte, Rezensionen, Literaturkritik, Wirkung, historische Rezeptionsforschung, Korpusanalyse, Sentimentanalyse, Zeitschriften, Periodika, 19. Jahrhundert

Keywords

literary history, reviews, literary criticism, effect, historical reception studies, corpus analysis, Sentiment analysis, journals, periodicals, 19th century

Inhaltsverzeichnis

1 Einleitung

Die Anzahl an Retrodigitalisaten historischer Literatur- und Kulturzeitschriften ist in den letzten Jahren stark angewachsen. Zwar ist die Volltexterkennung der zumeist in Fraktur gedruckten Texte im Augenblick noch ungenügend, doch lässt sich erwarten, dass es in wenigen Jahren möglich sein wird, repräsentative Korpora der Zeitschriftenkommunikation vergangener Epochen aufzubauen und mit digitalen Analysewerkzeugen auszuwerten. Dies wäre ein großer Gewinn für die Literaturgeschichtsschreibung, die sich bezüglich der Rezeption literarischer Werke bislang nur auf engbegrenzte Fallstudien stützen kann. Die literaturgeschichtliche Rezeptionsanalyse (Mellmann/Willand 2013) interessiert sich unter anderem dafür, wie literarische Werke vom zeitgenössischen Publikum bewertet wurden, d.h. welche Werke als ästhetisch hochwertig und welche als künstlerisch misslungen angesehen wurden.[1] Ein digitales Analyseinstrument, das auf einen bestimmten Redegegenstand bezogene Wertungen abzugreifen ermöglicht, liegt mit der sogenannten Sentimentanalyse (engl. „sentiment analysis" oder „opinion mining") vor. Eine Anwendung dieses Instruments auf das literaturkritische Schrifttum vergangener Epochen, das sich zu einem großen Anteil in Zeitschriften befindet, sollte also auch solche historischen Wertungsprozesse sichtbar machen können. Ein geeignetes Verfahren für diesen Zweck zu entwickeln, war die Aufgabe, die wir uns in unserem Projekt „SentiLitKrit_19-II"[2] gestellt haben.

Als Untersuchungszeitraum haben wir die zweite Hälfte des 19. Jahrhunderts gewählt, da Zeitschriften in diesem Zeitraum eine besonders wichtige Rolle in der literarischen Kommunikation spielen und – im Vergleich zu dem diversifizierten Zeitschriftenangebot im frühen 20. Jahrhundert – noch eine relativ homogene Öffentlichkeit bilden. Da im Moment noch kaum Volltexte zu den hier einschlägigen Zeitschriften vorliegen, haben wir uns aus einer literaturwissenschaftlichen Anthologie historischer Rezensionen (Kreuzer 2006) ein Probekorpus aufgebaut, an dem wir erste Analysedurchgänge vornehmen konnten.

Bei der Optimierung unserer Methoden sind wir nach der Maxime „je einfacher, desto besser" verfahren, das heißt: die Vorgänge sollten möglichst wenig komplex und möglichst voraussetzungsarm sein. Denn auch wenn in naher Zukunft mit einer größeren Menge OCR-bearbeiteter Retrodigitalisate zu rechnen ist, wird die Textqualität wegen der zum Teil komplizierten Seitenaufteilung (z.B. Spaltendruck, Kopfzeilen, Fußnoten), der vielen Eigennamen und anderer Schwierigkeiten wohl immer noch zu schlecht und die Textmenge zu groß sein, um in größerem Umfang Preprocessing-Verfahren anwenden zu können.

Wir haben daraus für's Erste die Konsequenz gezogen, nicht die Textqualität zu verbessern zu versuchen, sondern umgekehrt die Analyseinstrumente so grobmaschig zu halten, dass sie auch mit schlechtem, „schmutzigem", linguistisch unstrukturiertem Textmaterial gute Annäherungsergebnisse erbringen können. Die entsprechenden Vorentscheidungen, die wir getroffen haben, werden in Abschnitt 2 erläutert.

[1] Andere Fragestellungen, die im vorliegenden Beitrag nicht aufgegriffen werden, wären etwa die Frage nach dem selektiven Verständnis (z.B. von Handlungsverläufen oder anderen Textbedeutungsaspekten), das sich in konkreten Rezeptionszeugnissen abzeichnet, oder die Frage nach den werkexternen Kontextualisierungen, die einzelne Rezipienten vorgenommen haben (z.B. mit pädagogischen, sozialen, ethischen Problemen der Zeit). Wir beschränken uns hier auf die Frage nach der ästhetischen Wertung, da diese sich für eine digitale Analyse am ehesten anbietet.

[2] Im Rahmen des literaturwissenschaftlichen Dilthey-Projektes „Historische Rezeptionsforschung" gefördert durch die Volkswagenstiftung.

In Abschnitt 3 werden die Ergebnisse von zwei Testläufen dargestellt, in denen wir zusätzlich zur Basismethode der bloßen Sentimentwort-Zählung außerdem eine automatische Klassifikation vorgenommen haben, bei der zunächst nur Sentimentwort und Polaritätswert, im zweiten Testlauf dann auch weitere Faktoren (POS, Tf-idf-Gewichtung, Wortvektoren) als Feature eingesetzt wurden.

2 Methodische Überlegungen

Lexikon-basierte Sentimentanalyse: Eine der häufig eingesetzten Methoden für Sentimentanalyse ist die lexikon-basierte Sentimentanalyse. Lexikon-basierte Sentimentanalyse erfordert ein Sentimentwort-Lexikon und eine Aggregationsmethode. Das Sentimentwort-Lexikon speichert die Indikatoren der Polarität (Sentimentwörter); die Aggregationsmethode aggregiert die Polarität der Indikatoren, die in einem Text vorkommen. Neben der lexikon-basierten Methode besteht außerdem die Möglichkeit, maschinelles Lernen einzusetzen. Features wie z.B. Wort oder Dependenzbaum können für Sentiment-Klassifikation verwendet werden. Text wird dann durch ein bag-of-words-Modell oder durch ein Dependenzbaum-Modell repräsentiert. Basiert auf die Häufigkeit der Wörter oder Teilbäume wird Sentiment-Klassifikation durchgeführt. Laut bisheriger Forschung ist dependenzbaum-basierte Sentiment-Klassifikation besonders gut geeignet, um die Polarität kurzer Texte zu analysieren (Kudo & Matsumoto 2004, Nakagawa et al., 2010). Es gibt einige deutsche Dependency-Parser wie z. B. Mate Tools[3] und Stanford Parser[4]. Ein Problem besteht darin, dass diese Tools auf digitale Korpora deutschsprachiger Texte aus dem 20. Jahrhundert trainiert wurden (z.B. wurde Mate Tools auf dem TIGER Korpus[5] trainiert). Ob sie sich auch für Texte des 19. Jahrhunderts eignen und in Digitalisaten von der beschriebenen schlechten Qualität funktionieren, ist unklar. Aus diesem Grund sind wir zunächst bei der simpleren einzelwort-basierten Methode geblieben.

Sentimentwort-Lexikon: Der wichtigste Indikator der Polarität sind sogenannte Sentimentwörter (engl. „sentiment words" oder „opinion words"). Positive Sentimentwörter sind zum Beispiel „gut", „perfekt", „wunderbar", negative „schlecht", „unfair", „schädlich". Eine Liste derartiger Wörter wird als Sentimentwort-Lexikon bezeichnet. Im Lexikon wird außerdem der Sentiment-Wert jedes Wortes gespeichert. Bei der Analyse eines Textes werden alle in einem Text vorkommenden Sentimentwörter herausgelesen; der Sentiment-Wert des Textes ist dann die Aggregation der Sentiment-Werte der gefundenen Sentimentwörter. Dieser Gesamt-Sentiment-Wert kann anschließend als Feature der Klassifikation eingesetzt werden. Da die Leistungsfähigkeit eines Sentimentwort-Lexikons von Faktoren wie Domäne, Zeit und Sprachregister beeinflusst wird, haben wir für die Untersuchung der literarischen Rezensionen des 19. Jahrhunderts ein eigenes Sentimentwort-Lexikon erstellt. Die Ausgangsbasis des Lexikons war die Sentimentwortliste des Projekts „SentiWS" (Remus et al. 2010)[6]. Diese Liste enthält sowohl Adjektive und Adverbien als auch Nomen und Verben. Der Sentiment-Wert der einzelnen positiven und negativen Wörter ist als Polarität im Intervall [-1; 1] angegeben. Da unklar ist, ob die Polarität der Wörter im gegenwärtigen Deutsch und in unserem Textkorpus übereinstimmen, und da wir weitere Sentimentwörter hinzufügen wollten, für die wir keinen ermittelten Polaritätswert haben, haben wir die Polaritätsangaben

[3] http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/matetools.html
[4] https://nlp.stanford.edu/software/lex-parser.shtml
[5] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html
[6] http://wortschatz.uni-leipzig.de/de/download

des Lexikons nicht übernommen, sondern den Grundwert jedes Sentimentworts mit 1 angesetzt. In einem ersten Bearbeitungsdurchgang wurde die Liste außerdem um für unser Korpus unbrauchbare oder überflüssige Wörter (wie z.B. „geil", „Konkurs", „deinstallieren") gekürzt und um relevante Wörter (wie z.B. „philiströs", „Kunstwert", „ergreifen") erweitert.

Wahl der Analyseeinheit: Da wir die Bewertung einzelner Autoren bzw. Werke[7] erheben wollen, wäre die zunächst sich empfehlende Analyseeinheit die des vollständigen Rezensionstextes. Doch sind damit zwei Probleme verbunden, die uns von dieser Lösung haben Abstand nehmen lassen: Erstens würde die Extraktion einzelner Rezensionen aus einem mehrtausendseitigem Zeitschriftenjahrgang einen erheblichen Aufwand bedeuten (noch mehr die Extraktion einzelner, einem bestimmten Werk gewidmeter Teile aus häufig auftauchenden Sammelrezensionen). Zweitens variieren die Rezensionen stark in der Länge und können je nach Fall lange Inhaltsangaben oder auch literaturgeschichtliche Exkurse (mit Erwähnung zahlreicher weiterer literarischer Autoren) enthalten, durch welche ggf. ein hohes Maß an sentiment-relevantem Wortschatz in dem Text auftaucht, der mit der Bewertung des rezensierten Autors in keiner Beziehung steht. Wir haben uns daher entschieden, 1-Satz-Textausschnitte zu extrahieren von Sätzen, in denen ein Autorname erwähnt wird. Zwar variieren auch diese Sätze stark in der Länge, aber man kann voraussetzen, dass die Satz-Einheit in den meisten der Fälle auch ungefähr eine Aussage-Einheit darstellt und so die meisten der in dem Satz auftauchenden Sentimentwörter in Bezug auf den besprochenen Autor stehen.[8] Ein Nachteil dieser Methode ist, dass Sätze mit Proformen („er", „der Dichter", „der Künstler") durch sie nicht erfasst werden. Es gibt keinen Grund zu der Annahme, dass klar wertende Stellungnahmen relevant häufiger mit expliziter Namensnennung einhergehen als mit Proformen, so dass eine Extraktion der Sätze mit expliziter Nennung nur eine letztlich unbegründete Teilmenge der potentiell relevanten Textausschnitte erfasst. Eine Einbeziehung der Proformen-Sätze würde aber wiederum ein linguistisches Preprocessing der Texte notwendig machen, wie wir es aufgrund der schlechten OCR-Qualität von Zeitschriften-Retrodigitalisaten vermeiden wollen. So haben wir uns mit der unvollkommenen Methode der Extraktion expliziter Erwähnungen begnügt. Wenn jede Rezension mindestens einen wertenden Satz mit expliziter Namensnennung enthält, der die Tendenz angibt, dürften sich – da die Ergebnisse am Ende pro Autor und Zeitabschnitt gemittelt werden – immer noch befriedigende Annäherungsergebnisse erzielen lassen, auch wenn nicht alle Bezugnahmen auf den Autor eines Textes darin eingegangen sind.

Aggregation: Die Aggregation der Sentiment-Werte geschieht normalerweise über schlichte Summierung der Werte der in einem Textausschnitt aufgefundenen Sentimentwörter. Für unsere Untersuchung aber haben wir eine komplexere Aggregationsfunktion (Abb. 1) für die Berechnung des Sentiment-Werts verwendet, da in unserem Korpus bisweilen mehrere Autoren in einem Satz genannt und unterschiedlich bewertet werden. Deshalb wurde die Entfernung zwischen Autornennung und Sentimentwort in die

[7] In den hier dargestellten Probeläufen haben wir uns auf Autornamen als Bezugspunkt beschränkt. Dasselbe Verfahren ließe sich aber prinzipiell statt mit Autornamen auch mit Werktiteln als Suchwort durchführen. Allerdings stellt sich das Problem nichtspezifischer Ausdrücke, die falsche Textstellen identifizieren (Autornamen wie „Holz" oder „Hauptmann"), hier noch stärker (Werktitel wie „Jugend" oder „Kinder der Welt"). Aber auch auf einigermaßen spezifische Gattungsnamen wie etwa „Dorfgeschichte" ließe sich die Methode anwenden.

[8] Alternativ wäre an 3-Satz-Ausschnitte zu denken (da die eigentliche Aussage in manchen Fällen einen Satz vorher oder nachher ausgedrückt ist und der Autorerwähnungssatz nur die Konkretisierung vornimmt), oder auch an eine stabile 100-Wort-Textumgebung oder Ähnliches. Vergleichende Testläufe mit unterschiedlich gewählten Analyseeinheiten stehen noch aus.

Aggregationsfunktion miteinbezogen: „se_j is a sentiment expression in sentence s, $dist(se_j, a_i)$ is the word distance between aspect a_i and sentiment expression se_j in s, and $se_j.ss$ is the sentiment score of se_j" (Liu 2015).

$$\text{score}\left(a_i, s\right) = \sum_{\text{se}_j \in s} \frac{\text{se}_j.\text{ss}}{\text{dist}\left(\text{se}_j, a_i\right)}$$

Als Anschauungsbeispiel mag der folgende Satz aus unserem Untersuchungskorpus dienen:

> Diese *Feinheit* und *Anmut* der Form, das *warme*, *stimmungsvolle Kolorit*, das die Werke **Daudet's** gleichsam beseelt und ein geheimnisvolles Band der *Sympathie* knüpft zwischen dem *Künstler* und seinem Publikum; dieses undefinirbare Etwas, das **Zola** selbst „l'expression personelle" nennt und das wir wohl am besten durch „individuelles Leben" übersetzen, ***fehlt* Zola** gänzlich.

Im Satz tauchen sieben positive Sentimentwörter (kursiv geschrieben), ein negatives Sentimentwort (fett und kursiv geschrieben) und drei Autornennungen (fett geschrieben) auf. Durch die einfache Summierung der Sentiment-Werte (7 – 1 = 6) würde der Satzes als positiv identifiziert. Aber durch den Einsatz der aspektgebundenen Aggregationsfunktion ist der Sentimentwert von „Daudet" 0,968, während der Sentimentwert von „Zola" -0,481 ergibt. (Der Wert des ersten „Zola" ist 0,326 und der des zweiten -0,806. Die zwei Werte werden summiert.)

Unigramme / N-Gramme: Sentiment-Lexika wie z. B. SentiWS enthalten normalerweise nur Einzelwörter (Unigramme), die positive und negative Polarität tragen. Das Vorkommen oder die Frequenz dieser Unigramme bildet die Basis der Sentimentanalyse. Neben den unigramm-basierten gibt es jedoch auch Analyseansätze, die N-Gramme (Bigramme, Trigramme etc.) verwenden. Die Frage, ob N-Gramme für die Sentimentanalyse besser geeignet sind als Unigramme, ist allerdings noch nicht geklärt. In einer Untersuchung mit Filmkritiken erwiesen sich Unigramme als das bessere Feature für die Sentiment-Klassifikation (Pang et al. 2002). In einer anderen Untersuchung, die mit Produkt-Reviews arbeitete, zeigte sich hingegen, dass Bigramme und Trigramme in manchen Fällen besser sind (Dave et al. 2003). Im Sinne der Einfachheit wollten wir zunächst einmal testen, wie gut Unigramme funktionieren. Allerdings sind wir in unseren Testläufen auf Fälle gestoßen, in denen die Verwendung von N-Grammen hilfreich sein könnte.[9]

Part of Speech: Unter part-of-speech (POS, word classes, morphological classes, oder lexical tags) versteht man die Klasse von Wörtern. Durch die Kennzeichnung der Wortarten werden die lexikalischen und grammatischen Informationen des Wortes berücksichtigt (vgl. Jurafsky/Martin 2009, S. 299). Im Kontext der Sentimentanalyse wird POS oft eingesetzt, um Ambiguitätsprobleme zu lösen. Ein Beispiel:

[9] Hierbei handelt es sich vor allem um idiomatische Kollokationen. Zum Beispiel kommt das Wort „deutsch" insgesamt zu häufig in neutraler Verwendung vor, um es in die Positiv-Wortliste aufzunehmen; aber es kommt in relevantem Umfang auch in Kollokationen wie „echt/wahrhaft deutsch" vor und stellt in diesen Kontexten eine starke Positivwertung dar. Deshalb haben wir im zweiten Versuch auch eine bigramm-basierte Klassifikation ausprobiert.

(1) Wann wirst du entdecken, daß du das **Gute** nicht zuerst von den andern verlangen darfst, sondern selber anfangen musst?

(2) **Gute** Nachricht für den Sparer: eine Aufwertung der Reichsmark-Guthaben ist in Sicht.

Das „Gute" in Satz (1) ist ein Nomen, das kein Sentiment des Sprechers ausdrückt. In Satz (2) hingegen ist „Gute" trotz Großschreibung ein Adjektiv und besitzt positiven Sentiment-Wert. Deshalb haben wir entschieden zu testen, ob das POS für unsere Sentiment-Klassifikation sinnvoll ist.

Tf-idf-Gewichtung: Im Information Retrieval wird das Tf-idf-Maß eingesetzt, um die Termgewichte in einer Textsammlung zu beurteilen. Die Termhäufigkeit tf (term frequency) zeigt an, wie wichtig ein Term innerhalb eines Dokuments ist. Die invertierte Dokumenthäufigkeit idf (inverse document frequency) repräsentiert, wie gut Dokumente anhand des jeweiligen Terms unterschieden werden können. Wenn ein Wort (Term) häufig in positiven (oder negativen) Sätzen, aber selten im Rest des Korpus vorkommt, wird dem Wort ein hohes Gewicht beigemessen. Wörter mit hohem Gewicht können als Unterscheidungsmerkmal eingesetzt werden, um positive und negative Sätzen zu unterscheiden. Deswegen ist es sinnvoll, diese Methode für unsere Analyse auszuprobieren.

Wortvektoren: Eine Möglichkeit, semantische Relationen mitzuberücksichtigen, sind sogenannte Word-Embeddings. Word-Embeddings bilden Wörter auf einen Vektorraum ab, in dem jedes Wort durch einen Vektor repräsentiert wird. Je nach Modell hat ein Vektor üblicher Weise zwischen 100 und 300 Dimensionen. Diese numerische Repräsentation drückt die kontextuelle Bedeutung eines Wortes aus. Word-Embeddings können zum Beispiel eingesetzt werden, um die semantische Ähnlichkeit von Wörtern zu messen. Ein bekanntes Beispiel ist: Vektor(woman) + Vektor(king) – Vektor(man) = Vektor(queen). Außerdem lassen Word-Embeddings sich zu Sentence-Embeddings erweitern, sodass auch ganze Sätze in Form von Vektoren repräsentiert werden können. Der Satzvektor entspricht dem Durchschnitt aller Vektoren der Wörter in dem Satz. Die Satzvektoren können dann als Feature für die Sentiment-Klassifikation verwendet werden (Rudkowsky et al. 2018). Da unsere Datenmenge nicht ausreichte, ein eigenes Modell zu trainieren, haben wir mit einem vortrainierten fastText-Modell[10] (Grave et al. 2018) gearbeitet.

Wahl des Klassifikationsalgorithmus: Die Daten sollen mittels Machine-Learning-Algorithmen wie z.B. SVM, K-nearest-neighbors, Random Forest untersucht und klassifiziert werden. Die Ergebnisqualität einer Klassifikation kann je nach gewähltem Algorithmus recht unterschiedlich ausfallen. Für die Wahl des bestgeeigneten Algorithmus bietet die Scikit-learn-Bibliothek[11] eine Anleitung (scikit-learn algorithm cheat sheet[12]) an. Unsere Daten sind manuell gelabelt und die Datenmenge ist kleiner als 100 Tausend. Mit Rücksicht auf diese Kriterien verwenden wir Linear SVC (Support Vector Classification).

[10] https://github.com/facebookresearch/fastText/blob/master/docs/crawl-vectors.md
[11] http://scikit-learn.org/stable/index.html
[12] http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

3 Untersuchung

3.1 Korpusaufbau

Das Fernziel für eine tatsächliche Untersuchung wäre die Erstellung eines Korpus aus 5-10 repräsentativen Zeitschriften über ca. 3 Jahrzehnte. Von einer solchen Korpuserstellung sind wir im Augenblick noch weit entfernt, aber bevor eine solche Untersuchung durchgeführt werden kann, muss ohnehin zuerst ein brauchbares Analyse-Instrument entwickelt werden. Zu diesem Zweck haben wir ein Probekorpus erstellt, an dem wir erste Versuche unternehmen und unsere Methoden optimieren können. Das Probekorpus besteht aus Rezensionen, die für den Zeitraum 1870-1889 in eine literaturwissenschaftliche Anthologie deutschsprachiger Literaturkritik (Kreuzer 2006, Bd. 1-2) aufgenommen wurden. Die Rezensionen wurden aus dieser modernen Ausgabe gescannt und durch OCR (optical character recognition)[13] in Textdaten umgewandelt. Fehler, die die automatische Texterkennung produziert hat (z.B. „rn" statt „m„) wurden in einer vollständigen Korrekturlesung beseitigt. Zum Ausgleich von Deklination und historischer Orthographie wurden die digitalisierten Textdaten normalisiert und lemmatisiert. Diese Aufgabe wurde durch das „DTA::CAB Web Service" (Jurish 2012)[14] erledigt. Anhand des Namensregisters der Anthologie wurden anschließend alle Sätze extrahiert, in denen Namen literarischer Autoren vorkommen. Dies ergab insgesamt 1731 Sätze.

3.2 Versuch 1: 48% Trefferquote der Basismethode und Erweiterung um automatische Klassifikation

Das Ziel des ersten Testlaufs lag vor allem bei der Optimierung und Anpassung der Sentimentwortliste. Zuerst wurden die 1-Satz-Textsnippets in einem manuellen Rating als positiv, neutral oder negativ annotiert. Diesem ersten Labeling zufolge bestand das Probekorpus aus 505 positiven, 909 neutralen und 317 negativen Sätzen. Für jeden Satz wurde das Vorkommen der Sentimentwörter unserer angepassten Liste gezählt und durch die Aggregationsfunktion ein Sentiment-Wert berechnet. War der Sentiment-Wert größer als 0, wurde der Satz als positiv gerechnet, war er kleiner als 0, als negativ. War der Wert gleich 0, galt der Satz als neutral. Das Testergebnis in Tabelle 1 zeigt einen Positiv-Bias: Während insgesamt nur 47,4% der Sätze richtig erkannt wurden, lag die Trefferquote (engl. „recall„) bei den als positiv annotierten Sätzen bei 77%. Die Genauigkeit (engl. „precision") der Erkennung lag allerdings nur bei 40%, d.h. es wurden viele Sätze fälschlicherweise als positive Sätze identifiziert. Von den neutralen und negativen Sätzen wurden jeweils nur 35% richtig erkannt. Die Erkennung der negativen Sätze stellte sich in diesem Test als besonders problematisch dar: sowohl die Genauigkeit als auch die Trefferquote liegen bei nur rund einem Drittel.

[13] Für die Texterkennung wurde die OCR-Software ABBYY FineReader 12 verwendet.
[14] http://www.deutschestextarchiv.de/demo/cab/

Tabelle 1: Ergebnis des ersten Testlaufs

Dimension	*richtig identifizierte Sätze*	*manuelle Annotation*	*Precision*	*Recall*	*F1 score*
Positiv	387	505	0,4	0,77	0,52
Neutral	322	909	0,7	0,35	0,47
Negativ	112	317	0,33	0,35	0,34

Angesichts dieses unbefriedigenden Ergebnisses haben wir einen ausführlichen Optimierungsdurchgang an unserem Sentimentwort-Lexikon unternommen. Basis dieser manuellen Überprüfung war eine Ergebnisdarstellung, die den analysierten Satz, die darin aufgefundenen Sentimentwörter sowie den aggregierten Sentiment-Wert zeigte. Sentiment-Wörter, die sich zu häufig als dysfunktional erwiesen, wurden aus dem Lexikon entfernt, und funktional erscheinende Wörter, die sich noch nicht auf der Sentimentwort-Liste befanden, wurden hinzugefügt. So wird etwa das Wort „gut" in unserem Korpus so häufig redensartlich gebraucht, dass es als positives Sentimentwort unbrauchbar ist; so zum Beispiel im folgenden Satz:

> Aus diesem Zusammenhang heraus erhob sich im Herbst 1889 die Persönlichkeit Gerhart Hauptmann's, der vorher so gut wie unbekannt gewesen war.
>
> (0,25, ['gut'], [])

Das Wort „Persönlichkeit" hingegen, das im 19. Jahrhundert eine starke Emphatisierung erfahren hat, wurde in die Liste positiver Sentimentwörter aufgenommen.

Außerdem ist es offensichtlich zu restriktiv, die Klassengrenze zwischen positiv, neutral und negativ auf 0 einzustellen. Wir haben deshalb eine automatische Klassifikation ausprobiert. Dazu wurden die Daten zuerst in 80% Trainingsdaten und 20% Testdaten aufgesplittet. Ein Support Vector Machine (SVM) Modell wurde auf die Trainingsdaten trainiert. Dabei wurden der Sentiment-Wert eines Satzes und das Vorkommen der Sentimentwörter im Satz als Feature benutzt. Zur Bewertung des Modells wurde eine 10-fache Kreuzvalidierung (Cross-Validation) durchgeführt. Der Anteil der richtig erkannten Sätze hat sich dadurch von 48% auf durchschnittlich 57% (+/- 7%) um knapp 10% erhöht. Bei einer Klassifikation nur zwischen positiven und negativen Sätzen betrug die durchschnittliche Trefferquote sogar 68% (+/- 10%).

3.3 Versuch 2: 66% Trefferquote der Basismethode und Evaluation weiterer Features für die automatische Klassifikation

Der zweite Testlauf wurde mit der überarbeiteten Sentimentwortliste durchgeführt. Außerdem wurden die 1-Satz-Textsnippets vorher einem erneuten manuellen Rating unterzogen, da sich die Gruppe der zuvor als neutral kategorisierten Sätze als zu heterogen erwiesen hat. In dem neuen Labeling wurden nur noch Sätze ohne jede Wertungsinformation (z.B. „Ebers' neuester Roman führt den Titel ‚Die Schwestern'.") als neutral kategorisiert; für Sätze mit sowohl positivem als auch negativem Wertungsgehalt („Vorzüge wie Fehler Daudet's sind im Ganzen die der realistischen Schule."), die zuvor als einander sozusagen ausgleichend, also neutral bewertet worden waren, wurde hingegen eine eigene Gruppe gebildet. Auch wurden zwei neue Gruppen für Sätze gebildet, für die sowohl eine neutrale als auch eine positive bzw.

negative Zuordnung durch die Computeranalyse zufriedenstellend erschiene,[15] sowie zwei Gruppen für Problemfälle.[16] Die Annotation wurde also von drei (positiv/neutral/negativ) auf acht Kategorien erweitert: (1) Pos, (2) Neg, (3) Neutr, (4) PosNeg, (5) Neutr-Pos, (6) Neutr-Neg, (7) Pos-Artefakte, (8) Neg-Artefakte.[17] Für die Untersuchung haben wir ausschließlich die 1687 *eindeutig* positiven, negativen und neutralen Sätze (Gruppen 1-3) verwendet.

In diesem Teilkorpus wurden insgesamt 65,6% der Sätze richtig erkannt, wenn die Klassengrenze auf 0 gesetzt wird (Tab. 2). Im Vergleich zur ersten Untersuchung ließ sich die absolute Zahl der richtig erkannten Sätze von 821 auf 1107 steigern, die Erkennung der negativen Sätze stellt sich jedoch immer noch als äußerst schwierig dar. Unser Ergebnis bestätigt die Beobachtung, dass sich bei lexikon-basierten Sentimentanalysen generell ein Positiv-Bias einstellt (Kennedy & Inkpen 2006).

Tabelle 2: Ergebnis des zweiten Testlaufs

Dimension	*richtig identifizierte Sätze*	*manuelle Annotation*	*Precision*	*Recall*	*F1 score*
Positiv	599	718	0,65	0,83	0,73
Neutral	375	677	0,78	0,55	0,65
Negativ	133	292	0,47	0,46	0,46

Bei der automatischen Klassifikation durch SVM betrug die durchschnittliche Trefferquote 64% (+/- 7%); und eine automatische Klassifikation nur zwischen positiven und negativen Sätzen ergab eine Trefferquote von 76% (+/- 7%). Wir haben uns von da an nur noch auf die positiven und negativen Sätze[18] sowie auf die Einbeziehung weiterer Features konzentriert. Sowohl die Lemmata, die Tokens, die Lemmata-Bigramme, die Wortvektoren, als auch POS, POS-Bigramme und der Sentiment-Wert jedes Satzes wurden kombiniert, um einen SVM-Klassifikator zu trainieren, eine bestmögliche Unterscheidung von positiven und negativen Sätzen vorzunehmen. Die Tests erfolgten als 10-fache Kreuzvalidierung. In Abb. 1 ist die Einzeltrefferquote-Verteilung der 10 Einzeldurchläufe von jeweiligen Features bzw. Feature-Kombinationen dargestellt.[19] Die besten Ergebnisse lagen bei einer durchschnittlichen Trefferquote von 80% (Test 6, 7 und 21).

[15] D.h. tendenziell wertungsfreie Sätze, die aber von einem grundlegend wohlwollenden bzw. kritischen Duktus getragen sind; z. B. „Während dieser zwanzig Jahre hat Anzengruber alle Bitterkeiten und alle Freuden des berühmten Mannes, des gekrönten und bekämpften Dichters in so reicher Fülle erlebt, wie nicht leicht ein andrer." (Neutr-Pos); „Die Stücke von Wilbrandt und Lindau sind von ihrem Publikum mit dem üblichen Applaus aufgenommen worden." (Neutr-Neg).

[16] Hierbei handelt es sich in der Hauptsache um Sätze mit mehreren Aussageebenen (z.B. „Wer jedoch diesen Andeutungen entnimmt, Liliencron sei ein ‚Tendenzdichter', erfüllt von Tagesmeinungen, von ‚zufälligen Wahrheiten', wie Spinoza sich ausdrücken würde, der hat mich falsch verstanden.") oder anderweitig komplexe Formulierungen, die Artefakte erwarten lassen (z.B. „Für angenehmen Zeitvertreib sind Mackays Bücher nicht berechnet." D.h.: sie sind anspruchsvoller als kommerzielle Konsumware).

[17] Die Daten sind hinterlegt auf http://github.com/dkltimon/SentiLitKrit_19-II.

[18] Der Untersuchung von Pang et al. 2002 und Taboada et al, 2010 folgend wurden die neutralen Sätze ignoriert, um zu ermöglichen, dass der Klassifikator sich besser auf die positive und negative Polarität konzentriert.

[19] In der Abbildung steht „Le/Bi" bzw. „POS/Bi" für Lemmata-Bigramme bzw. POS-Bigramme.

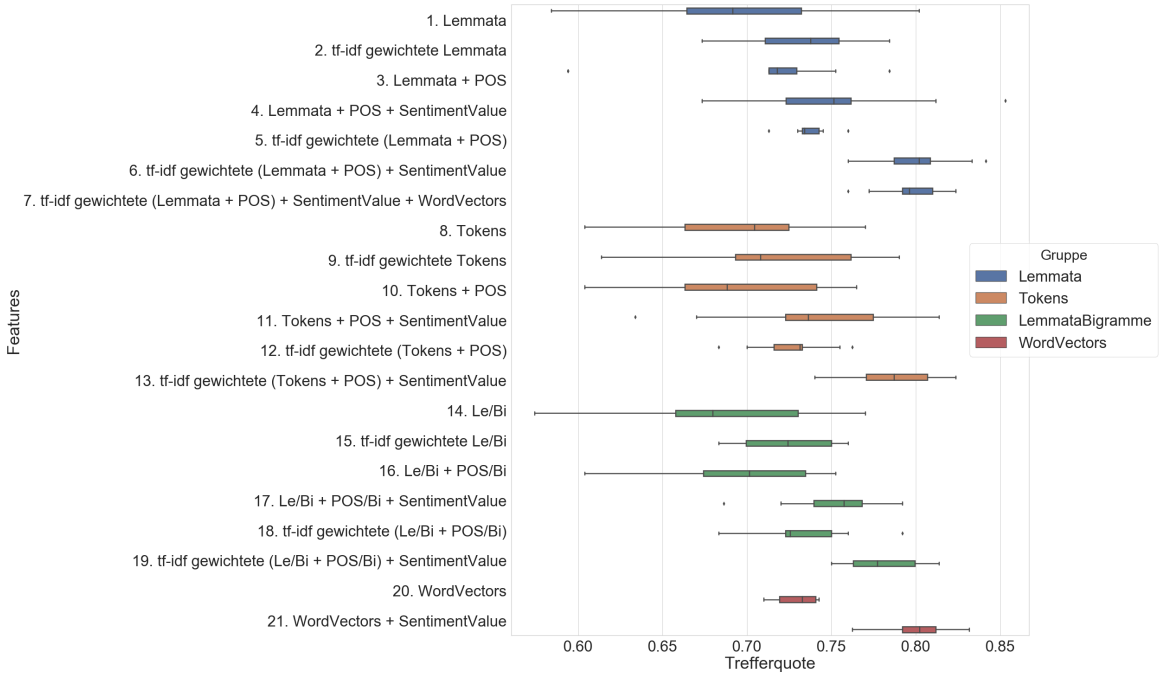

Abbildung 1: Einzeltrefferquote-Verteilung der 10 Einzeldurchläufe von Features und. Feature-Kombinationen

Der Unterschied zwischen den drei Gruppen (Lemmata: Test 1-7, Tokens: Test 8-13, Lemmata-Bigramme: Test 14-19) fällt relativ gering aus. Ob Lemmata, Tokens oder Lemmata-Bigramme als Feature der Klassifikation eingesetzt werden, scheint also nicht ausschlaggebend zu sein. Durch die Kombination mit weiteren Features aber lässt sich durchaus noch eine Effizienzsteigerung erzielen. Das POS-Feature schneidet dabei am schlechtesten ab: Für sich allein genommen verbessert es das Ergebnis nur um 1-2% (Test 3, 10 und 16). Die Trefferquote erhöht, wenn man die Features mit Tf-idf-Gewichtung verwendet (Test 2, 9 und 15 bzw. 5, 12 und 18). Durch Hinzunahme von Sentiment-Wert des Satzes (Test 4, 11 und 17) hingegen lässt sich die Trefferquote auch erhöhen; durch Kombination von beidem sogar noch mehr (Test 6, 13 und 19). Wir haben außerdem getestet, ob eine Kombination von allen Features (Frequenz von Lemmata, Tokens, Lemmata-Bigramme, POS, POS-Bigramme und Sentiment-Wert) noch effizienter sein könnte. Eine höhere durchschnittliche Trefferquote wurde dadurch aber nicht erreicht, sie lag bei 78%.

Auch Word-Embeddings brachten keine zusätzliche Verbesserung: Wenn man die Ergebnisse von Test 6 und 7 vergleicht, ist keine Erhöhung der Trefferquote zu beobachten, obwohl in Test 7 Wortvektoren als zusätzliches Feature hinzugenommen wurden. Dass die Einzeltrefferquoten von Test 5 und 12 ähnlich verteilt sind wie in Test 20 (und in Test 6 und 7 ähnlich wie in Test 21), lässt darauf schließen, dass Wortvektoren bei einer Klassifikationsaufgabe in etwa gleich gut funktionieren wie die Kombination aus tf-idf-gewichteten Lemmata- und POS-Frequenzen. Das liegt vermutlich daran, dass sie die gleiche (oder zumindest ähnliche) linguistische Informationen enthalten.

4 Resümee

Unser Ziel war die Anpassung gängiger Methoden der Sentimentanalyse für die Untersuchung literaturkritischen Schrifttums des späten 19. Jahrhunderts. Dieses Ziel ist noch nicht erreicht, aber nach zwei Versuchsläufen lässt sich ein erstes Zwischenfazit ziehen.

Die Domänenanpassung der Sentimentwortliste ist als ein work-in-process aufzufassen: Eine grobe Anpassung an das historisch spezifische Textgenre „Literaturkritik" ist in zwei Schritten vollzogen worden; weitere Feinjustierungen durch eine manuelle Ergebniskontrolle stehen noch aus.[20] Nach den ersten Erfahrungen lässt sich sagen, dass nicht nur explizit wertungstragende Wörter (wie z.B. „hervorragend", „genial", „dilettantisch") als Sentimentwörter in Frage kommen, sondern auch Wörter, die ein bestimmtes Sprachregister markieren (z.B. „Gestalt" für respektvolles vs. „Geschöpf" für ironisches bis despektierliches Sprechen), berücksichtigt werden müssen, ja in dieser vielfach von höflich-euphemistischem Stil geprägten Textsorte sogar bisweilen die ausschlaggebenderen sind.

Lässt man Schwierigkeiten wie Ironie, mehrere Aussageebenen (z.B. Zitate) und Understatement, wie sie in dem stark rhetorisierten Genre zuhauf auftreten, zunächst einmal außer Acht, stellt sich vor allem die Frage, wie mit Verneinungen umzugehen ist. Die hier aufgeführten Ergebnisse stammen aus Testläufen, in denen Negationswörter nicht eigens berücksichtigt wurden. Der Umgang mit Negationswörtern stellt in der lexikon-basierten Sentimentanalyse immer ein Problem dar. Eine einfache Möglichkeit der Einbeziehung besteht darin, den Wörtern zwischen einem Negationswort und dem nächsten Satzzeichen eine spezielle Markierung (z. B. „NOT_") zu geben und die Polarität der betroffenen Wörter einfach umzudrehen (Das & Chen, 2001). Diese Methode wurde ursprünglich für die Analyse englischer Texte eingesetzt. In unserem Probekorpus hat sie nicht durchgängig bewährt. Wir haben sowohl im ersten als auch im zweiten Versuch einen Testlauf mit und einen ohne Berücksichtigung von Negationswörtern unternommen und dabei keine signifikante Änderung des durchschnittlichen Testergebnisses beobachten können; funktionale und dysfunktionale[21] Effekte der Einbeziehung von Negationswörtern hielten sich offenbar die Waage. Eine bessere Lösung dieses Problems würde eine vertiefte syntaktische Analyse der deutschen Sätze erfordern, wie wir sie im Hinblick auf die schlechte Qualität der Textdaten vermeiden wollen.

Es hat sich bewährt, den ermittelten Polaritätswert als Feature einer automatischen Klassifikation einzusetzen und mit der Tf-idf-Gewichtung (Tf-idf) zu kombinieren. In dieser Kombination erreichten wir für Lemmata-Unigramme, Tokens und Lemmata-Bigramme jeweils die höchste Trefferquote. Dieses Ergebnis von rund 80% Trefferwahrscheinlichkeit basiert wohlgemerkt nur auf einer Teilmenge unseres Probekorpus: den 718 manuell als eindeutig positiv bewerteten und den 292 als eindeutig negativ bewerteten Textsnippets. Die Entscheidung, weniger eindeutige und offenkundige Problemfälle ab Versuch 2 nicht mehr mit einzubeziehen, wurde von der Überlegung getragen, dass die Methodenentwicklung angesichts der schlechten Ergebnisse im ersten Testlauf besser in zwei Schritten vollzogen werden sollte: Wir versuchen in einem ersten Schritt, an einem idealisierten Korpus ein möglichst treffsicheres Instrument zu entwickeln. Bevor dieses Instrument in einem empirischen Untersuchungskorpus eingesetzt werden kann, muss es erst in einem zweiten Schritt auf den jetzt nichtberücksichtigten Teil unseres

[20] Die letzte Fassung unserer Sentimentwort-Liste ist hinterlegt auf http://github.com/dkltimon/SentiLitKrit_19-II.

[21] Ein Problembeispiel ist der Satz: „Ein Buch von Theodor Storm braucht **nicht NOT_empfohlen** zu werden." Durch die Umdrehung der positiven Polarität von „empfehlen" bekommt dieser positive Satz einen negativen Sentiment-Wert.

Probekorpus angewandt und die Ergebnisse kontrolliert werden. Denn es ist nicht garantiert, dass ein für eindeutige Fälle gut funktionierendes Instrument auch für problematischere Fälle vergleichsweise bessere Ergebnisse liefert als ein für eindeutige Fälle weniger optimales Instrument. Dies ist eine empirische Frage, die erst im zweiten Schritt der Methodenentwicklung beantwortet werden kann. Gegenwärtig kann noch nicht einmal der erste Schritt als abgeschlossen gelten, auch wenn die Trefferwahrscheinlichkeit der automatischen Klassifikation in der genannten Kombination auf gute 80% erhöht werden konnte. Wünschenswert wäre neben der weiteren Optimierung des Sentimentwort-Lexikons zunächst noch eine vergleichende Evaluation unterschiedlicher Analyseeinheiten, ehe man zum zweiten Schritt übergeht. Auch an eine Erweiterung des Probekorpus um Textausschnitte aus einem bereits vorliegenden Zeitschriftendigitalisat[22] wäre zu denken, um einer möglichen Verzerrung des Probekorpus durch den Anthologiencharakter unserer Datenbasis entgegenzuwirken.

[22] Wir haben aus dem öffentlich zugänglichen Korpus *Die Grenzboten* (digitalisiert durch die Staats- und Universitätsbibliothek Bremen, http://brema.suub.uni-bremen.de/grenzboten) bereits 1-Satz-Textsnippets zu den in unserem Korpus besprochenen Autoren extrahiert. Die sehr aufwändige händische Selektion von Falschpositiven sowie das manuelle Rating stehen jedoch noch aus.

5 Literaturverzeichnis

Boucher, Jerry D. / Osgood, Charles E. (1969): The Pollyanna hypothesis. Journal of Verbal Learning and Verbal Behaviour, 8:1-8.

Das, Sanjiv / Chen, Mike (2001) Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proc. of the 8th Asia Pacific Finance Association Annual Conference (APFA 2001).

Dave, Kushal / Lawrence, Steve / Pennock, David M. (2003): Mining the peanut gallery: Opinion extractionand semantic classification of product reviews. In Proceedings of WWW, pages 519–528.

Grave, E. / Bojanowski, P. / Gupta, P. / Joulin, A. / Mikolov, T. (2018): Learning word vectors for 157 languages. arXiv preprint arXiv:1802.06893.

Jurafsky, D. / Martin, J. H. (2009): Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Prentice Hall series in artificial intelligence, 1-1024.

Jurish, B. (2012): Finite-state Canonicalization Techniques for Historical German. PhD thesis, Universität Potsdam (defended 2011).

Kennedy, Alistair / Inkpen, Diana (2006): Sentiment classification of movie and product reviews using contextual valence shifters. Computational Intelligence, 22(2):110–125.

Kreuzer, Helmut (Hg.) (2006): Deutschsprachige Literaturkritik 1870-1914. Eine Dokumentation. Unter Mitarbeit von Doris Rosenstein. 4 Bde. Frankfurt am Main: Lang.

Kudo, Taku / Matsumoto, Yuji (2004) A boosting algorithm for classification of semi-structured text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

Liu, Bing (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press.

Mellmann, Katja / Willand, Marcus (2013): Historische Rezeptionsanalyse. Zur Empirisierung von Textbedeutungen. In: P. Ajouri, K. Mellmann & C. Rauen (Hg.): Empirie in der Literaturwissenschaft. Münster: Mentis, S. 263–281.

Nakagawa, T. / Inui, K. / Kurohashi, S. (2010, June). Dependency tree-based sentiment classification using CRFs with hidden variables. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 786-794). Association for Computational Linguistics.

Pang, Bo / Lee, Lillian / Vaithyanathan, Shivakumar (2002): Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86.

Remus, Robert / Quasthoff, Uwe / Heyer, Gerhard (2010): SentiWS - a Publicly Available German-Language Resource for Sentiment Analysis. In: Proceedings of the 7th International Language Ressources and Evaluation, pages 1168-1171.

Rudkowsky, Elena / Haselmayer, Martin / Wastian, Matthias / Jenny, Marcelo / Emrich, Štefan / Sedlmair, Michael (2018): More than Bags of Words: Sentiment Analysis with Word Embeddings. In: Communication Methods and Measures, 12:2-3, 140-157, DOI: 10.1080/19312458.2018.1455817

Taboada, Maite / Brooke, Julian / Tofiloski, Milan / Voll, Kimberly / Stede, Manfred (2011): Lexicon-based methods for sentiment analysis. In: Computational linguistics, 37(2), 267-307.