

"Kann man denn auch nicht lachend sehr ernsthaft sein?" – Zum Einsatz von Sentiment Analyse-Verfahren für die quantitative Untersuchung von Lessings Dramen

Schmidt, Thomas

thomas.schmidt@sprachlit.uni-regensburg.de
Lehrstuhl für Medieninformatik, Universität Regensburg

Burghardt, Manuel

manuel.burghardt@ur.de
Lehrstuhl für Medieninformatik, Universität Regensburg

Dennerlein, Katrin

katrin.dennerlein@uni-wuerzburg.de
Institut für Deutsche Philologie, Julius-Maximilians-Universität Würzburg

Sentiment Analyse und Dramenanalyse

Sentiment Analyse (SA) beschreibt eine Reihe von computergestützten Methoden zur Prädiktion der Polarität eines Texts, versucht also vereinfacht gesagt automatisiert herauszufinden, ob ein Text ein positives oder negatives Gefühl ausdrückt (Liu 2016). Darüber hinaus werden teilweise auch komplexere emotionale Kategorien (wie z.B. Zorn und Freude) betrachtet (Mohammad & Turney 2010). Zentrale Anwendungsfelder der SA sind bislang vor allem die Analyse von Online-Reviews (McGlohan, Glance & Reiter 2010) und Social Media-Daten (Kouloumpis, Wilson & Moore 2011).

Zur Analyse von literarischen Texten mittels SA-Techniken finden sich bislang nur wenige Studien, z.B. zu Märchen (Alm, Roth & Sproat 2005) und Romanen (Kakkonen & Kakkonen 2011; Elsner 2012; Jannidis et al. 2016). Auf größeren Textkorpora wurde getestet, inwiefern SA-Werte eines Textes und Emotionskurven von Texten zur Genreklassifikation verwendet werden können (Kim, Padó & Klinger 2017) und wie begriffsgeschichtliche Bedeutungsverschiebungen in literarischen Texten mithilfe von erweiterten SA-Methoden erforscht werden können (Buechel, Hellrich & Hahn 2017). In Dramentexten hat man bisher die Verteilung von emotionalen

Kategorien (Mohammad 2011) oder die Entwicklung von Figurenbeziehungen (Nalisnick & Baird 2013) in Shakespeare-Dramen untersucht. Auch der vorliegende Beitrag beschäftigt sich mit dem Einsatz von SA im Bereich der Dramenanalyse. Es werden erstmals systematisch verschiedene Methoden der SA für Dramen getestet und evaluiert. Zudem wird exploriert, inwiefern bisher in der Literaturwissenschaft erforschte Aspekte von Dramen mithilfe der SA erfasst werden und inwiefern die SA auch für die Gewinnung neuer literaturwissenschaftlicher Erkenntnisse eingesetzt werden kann.

Das im Rahmen dieser Studie verwendete Lessing-Korpus umfasst ein mit Strukturinformationen annotiertes Dramenkorpus mit 11 Dramen, bestehend aus insgesamt 8224 Einzelrepliken. Sämtliche Dramen wurden über die Plattform *TextGrid*¹ bezogen, so dass alle im Rahmen dieses Beitrags entwickelten Tools auch auf andere *TextGrid*-Dramen anwendbar sind. Mit dem am besten evaluierten SA-Verfahren wurde eine webbasierte Anwendung zur Analyse und Visualisierung von Sentiment-Verteilungen und -Verläufen implementiert.

Evaluation unterschiedlicher SA-Verfahren

Lexikonsbasierte SA

Innerhalb der SA unterscheidet man zwei wesentliche Ansätze: (1) die Nutzung maschinellen Lernens und (2) die Verwendung lexikonbasierter Verfahren. Für das erstgenannte Vorgehen ist typischerweise ein mit Sentiment-Informationen annotiertes Trainingskorpus notwendig (D'Andrea et al. 2015), welches für die Dramenanalyse bislang nicht vorliegt. Aus diesem Grund werden in der vorliegenden Arbeit lexikonbasierte Verfahren eingesetzt. Ein Sentiment-Lexikon ist dabei eine Wortliste, in der für jedes Wort Sentiment-Informationen angegeben sind (Liu 2016: 10), also z.B. ob es positiv oder negativ konnotiert ist und in welchem Ausmaß (Polaritätsstärke). Ein derartiges Wort nennt man auch *sentiment bearing word* (SBW; Liu 2016: 189).

SA-Parameter

Folgende SA-Optionen wurden in unterschiedlichen Kombinationen systematisch evaluiert:

i) **Lexika** – Es wurden fünf zentrale Sentiment-Lexika für den deutschsprachigen Bereich herangezogen: *SentiWortschatz* (SentiWS; Remus, Quasthoff & Heyer 2010), die *Berlin Affective Word List – Reloaded* (Bawl-R; Vo et al. 2009), die deutsche Version des *NRC Emotion-Association Lexicon* (NRC, Mohammad & Turney 2010), ein Lexikon von Clematide & Klenner (2010; im folgenden CK genannt) und das *German Polarity Clues* (GPC; Waltinger 2010). SentiWS, Bawl-R und CK enthalten

Polaritäten und Polaritätsstärken, das NRC und GPC nur Polaritätsangaben. Das NRC enthält des Weiteren Annotationen zu acht unterschiedlichen Emotionen (Zorn, Furcht, Erwartung, Freude, Vertrauen, Ekel, Traurigkeit, Überraschung).

ii) Historisch-linguistische Varianten – Über ein Tool des Deutschen Text-Archivs von Jurish (2011) wurde die Option der Lexikon-Erweiterung mit historischen linguistischen Varianten der Originalwörter untersucht.

iii) Stoppwortlisten – Analog zu Saif et al. (2014) wurde der Einfluss der Verwendung von insgesamt drei unterschiedlichen Stoppwortlisten auf die Qualität der SA untersucht. Grund hierfür ist, dass durch verschiedene Kombination der Verfahren Sentiment-tragende Stoppwörter entstehen. Neben herkömmlichen Stoppwörtern wurden dabei auch Listen mit hochfrequenten Wörtern des Korpus untersucht. Dadurch wird der Einfluss von Wörtern analysiert, die zwar als sentiment-tragend in SA-Lexika ausgezeichnet werden, aber aufgrund der häufigen Nutzung im Korpus ein ungleichmäßiges Sentiment-Gewicht erzeugen (z.B. Herr, Fräulein).

iv) Lemmatisierung – Eine weitere untersuchte Verarbeitungsform für die SA ist die Lemmatisierung. Als Lemmatisierer werden der *Pattern-Lemmatisierer* (De Smedt & Daelemans 2012) der Python-Bibliothek *textblob* und der Python-Wrapper des *treetagger*-Tools (Schmid 1995) evaluiert. Viele SA-Lexika enthalten lediglich Grundformen. Aufgrund der Probleme und Schwierigkeiten der Lemmatisierung im Deutschen (Eger, Gleim & Mehler 2016) soll vergleichend untersucht werden, welcher Lemmatisierer die besten Ergebnisse in Kombination mit Lexika erzielt. Ferner enthalten einige SA-Lexika manuell angegebene flektierte Wortformen. Es wird somit auch die automatische Lemmatisierung mit der manuellen Erweiterung verglichen.

SA-Metriken

Alle nachfolgenden Berechnungen wurden bezüglich aller kombinatorischen Möglichkeiten der soeben beschriebenen SA-Parameter durchgeführt. Dabei werden die jeweiligen SA-Metriken nach Term-Zähl-Methodik (Kennedy & Inkpen 2006) berechnet, d.h. ein Text wird hinsichtlich vorhandener SBWs untersucht, positive und negative Wörter ausgezählt und für einen Polaritätswert die positive von der negativen Zahl subtrahiert. SA-Metriken wurden auf folgenden Ebenen über die jeweils zugehörigen Texte kalkuliert: Drama, Akte, Szenen, Repliken sowie Sprecher und Sprecherbeziehungen pro Drama, Akt, Szene und Replik. Die Beziehungen zwischen den Figuren wurden nach einer Heuristik von Nalisnick & Baird (2013) berechnet.

Erstellung des Gold Standards

Zur systematischen Evaluation der Prädiktionsleistung der verschiedenen SA-Ansätze wurde ein Evaluationskorpus bestehend aus 200 Repliken erstellt. Bei der Auswahl der Repliken wurde darauf geachtet, dass die dramenspezifische Verteilung berücksichtigt wird, längere Dramen sind also mit mehr Repliken vertreten. Ferner wurden nur solche Repliken aufgenommen, die mindestens 19 Wörter umfassen. Diese Länge entspricht etwa -25% des Mittelwerts des Gesamtkorpus und vermeidet damit die Selektion von zu kurzen Repliken. Es wurde insgesamt auf eine gleichmäßige Längenverteilung geachtet.

Die Repliken wurden von insgesamt fünf Personen (4 weiblich, 1 männlich; alle jeweils mit Deutsch als Muttersprache) jeweils unabhängig voneinander bezüglich deren Polaritätswirkung bewertet. Die Polarität jeder Replik wurde jeweils sechswertig (sehr negativ, negativ, neutral, gemischt, positiv, sehr positiv) und binär (positiv, negativ) bewertet. Die Annotationen wurden bezüglich des Übereinstimmungsgrades analysiert. Dazu wurden das Übereinstimmungsmaß Fleiss' Kappa (Fleiss 1971) sowie der Durchschnittswert der prozentualen Übereinstimmung aller Annotatoren und Annotatorinnen berechnet (vgl. Tabelle 1).

Annotationsskala	Fleiss' Kappa	Prozentuale Übereinstimmung
Polarität-sechswertig	0,22	40%
Polarität-binär	0,47	77%

Tabelle 1. Annotator agreement.

Man erkennt eine geringe Übereinstimmung für die Bewertungsskala mit sechsstufiger Polarität und eine moderate Übereinstimmung für die binäre Variante. Die Ergebnisse verhalten sich konform zu verwandten Studien bei der Interpretation literarischer Texte (Alm & Sproat 2005). Als finale Annotation für eine Replik wird die binäre Polarität gewählt, die die Mehrheit der Annotatoren und Annotatorinnen ausgewählt haben (Endresultat: 139 negativ, 61 positiv).

Evaluationsmaße

Als Evaluationsmaße wurden Genauigkeit (accuracy), Recall, Precision und F-Werte (Gonçalves et al. 2013) herangezogen. Abb. 1 zeigt einen Ausschnitt aus den je fünf besten Kombinationen pro Lexikon, geordnet nach Genauigkeit.²

Metric	DTAExtension	Lemmatization	Stopwords	accuracy	F-MeasureAvera
polaritySentiWS	dtExtended	textblob	noStopwordList	0,67	0,6373626374
polaritySentiWS	dtExtended	tokens	noStopwordList	0,665	0,5775402755
polaritySentiWS	dtExtended	treeTagger	noStopwordList	0,65	0,6042514699
polaritySentiWS	dtExtended	treeTagger	enhancedList	0,615	0,5558247527
polarityCd	dtExtended	treeTagger	enhancedList	0,595	0,5644107445
polarityCd	dtExtended	treeTagger	enhancedFilterex	0,585	0,5607419756
polarityCd	dtExtended	textblob	enhancedList	0,565	0,5556577032
polarityGpc	noExtension	textblob	enhancedFilterex	0,56	0,5397008055
polarityGpc	dtExtended	textblob	enhancedFilterex	0,56	0,5397008055
polarityCd	dtExtended	treeTagger	standardList	0,55	0,5499549955
polarityCdDichotom	dtExtended	treeTagger	enhancedList	0,535	0,5102040816
polarityCdDichotom	dtExtended	treeTagger	enhancedFilterex	0,53	0,5123975516
clearlyPolarityCombined	dtExtended	textblob	enhancedList	0,51	0,5028409091
clearlyPolarityCombined	dtExtended	treeTagger	enhancedList	0,505	0,4854336131
polaritySentiWSDichotom	dtExtended	tokens	noStopwordList	0,5	0,486863711

Abbildung 1: Ausschnitt aus der detaillierten Ergebnistabelle zur Evaluation der SA-Kombinationsmöglichkeiten.

Ergebnisse der Evaluation

Nachfolgend erfolgt eine überblicksartige Zusammenstellung einiger zentraler Ergebnisse aus der Evaluation:

- Eine explizite Lemmatisierung führt zu einer verbesserten Leistung. Beide Lemmatisierer erzielen dabei meist ähnliche Ergebnisse. Die Lexikonerweiterung durch historische Varianten macht die explizite Lemmatisierung jedoch weitestgehend unnötig, da hierbei auch eine grundlegende Lemmatisierung inkludiert ist.
- Es zeigt sich eine konsistente Verbesserung durch die Lexikonerweiterung mittels der Wort-Varianten aus dem Tool von Jurish (2011).
- Stoppwortlisten haben nur auf vereinzelte Lexika (GPC, CK) einen merklich positiven Einfluss.
- Lexika mit Polaritätsstärken sind meist besser als reine Term-Zähl-Verfahren desselben Lexikons.
- Das Lexikon, dass die höchsten Genauigkeiten für die SA erzielt, ist SentiWS
- Die beste Leistung (unter Analyse aller Metriken) erzielt das erweiterte SentiWS mit den Polaritätsstärken, lemmatisiert mittels Pattern-Lemmatisierer und ohne Stoppwortliste (Genauigkeit = 0,67; F-Wert = 0,64). Die Erkennungsrate ist besser als die random baseline von 0,576 aber schlechter als viele Erkennungsraten auf anderen Anwendungsgebieten der SA (Vinodhini & Chandrasekran 2012).

Aufgrund der Tatsache, dass hier ein verhältnismäßig simpler SA-Ansatz gewählt wurde und bereits menschliche Annotatoren und Annotatorinnen Schwierigkeiten mit der Polaritätsbestimmung haben, sind die Ergebnisse insgesamt durchaus positiv zu bewerten.

Online-Tool

Abschließend wurde auf Basis des besten SA-Ansatzes ein Web-Tool für die SA bei Dramen entwickelt. Dieses bietet interaktive Visualisierungen der

Sentiment-Verteilungen und -Verläufe für alle berechneten Ebenen. Neben den SentiWS-Metriken wurden auch die Emotionskategorien des NRC integriert. Über das Tool kann man erste Fallstudien auf Dramen-, Akt-, Szenen-, Repliken-, Sprecher- und Sprecherbeziehungsebene durchführen. Die SA-Komponente ist online verfügbar.³

Trotz der historischen Differenz stimmen die Ergebnisse der automatischen SA tendenziell mit dem überein, was man in der Dramengeschichte über Bewertungen von Figuren und deren Verhalten weiß. Zusätzlich ist aber ein wichtiger heuristischer Mehrwert zu beobachten: eine Analyse allein auf der Basis von Sentiment-Zuschreibungen führt dazu, dass man das Augenmerk gezielt auf Fakten des Textes richtet, die bisher nicht berücksichtigt wurden.

Im Folgenden einige Beispiele für die Bestätigung bekannter Ergebnisse und für Entscheidungen von Analysefragen:

Fallstudie: Minna von Barnhelm

Die Analyse von Minna von Barnhelm zeigt, dass die negativen emotionalen Bewertungen insgesamt gegenüber den positiven deutlich überwiegen (vgl. Abb. 2). Dieser Befund bestätigt die bekannte Erkenntnis, dass Lessing das Schema des rührenden Lustspiels verwendet hat. Während die Komik im Stück eher das Ergebnis von Schlussprozessen ist, geht es auf der wörtlichen Ebene überwiegend um ernste Vorwürfe und drohenden Identitäts- und Beziehungsverlust.

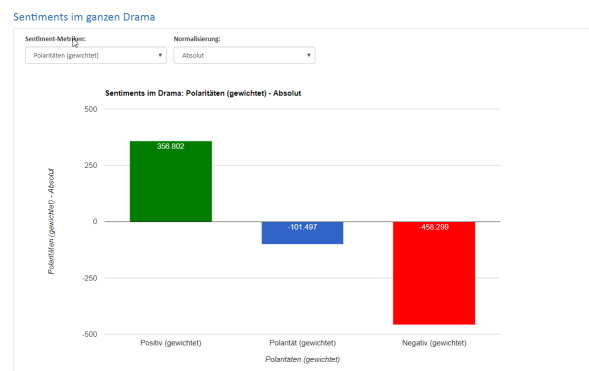


Abbildung 2: Polaritätsverteilung im Drama – Minna von Barnhelm

Es ist verschiedentlich behauptet worden (Saße 1993), Minna und nicht Tellheim sei die lächerliche Figur des Stücks. Die Sympathienlenkung auf der wörtlichen Ebene des Textes, die in der unten stehenden Sentimentverteilung pro Akt abgebildet ist, kann dazu herangezogen werden, diese Frage negativ zu bescheiden (vgl. Abb. 3). Es ist eine auffällige Abweichung der Polarität im zweiten Akt erkennbar. In diesem Akt tritt Minna von Barnhelm zum ersten Mal auf, Tellheim jedoch nicht.

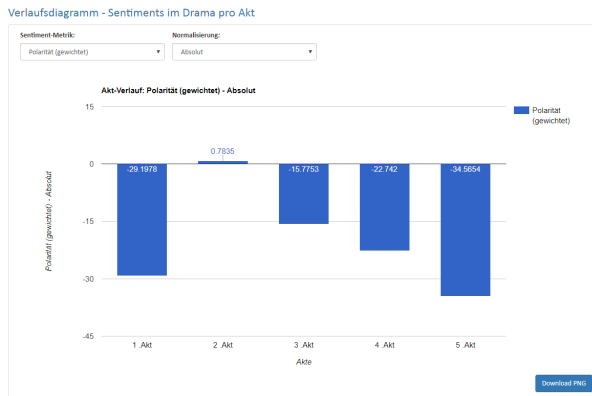


Abbildung 3: Polaritätsverlauf pro Akt – Minna von Barnhelm

Fallstudie: Emilia Galotti

Die letzte Visualisierung kann genutzt werden die Frage zu diskutieren, warum Emilia in Lessings Drama „Emilia Galotti“ sterben muss (vgl. Abb. 4). Auffällig ist hier die starke negative Bewertung Emilias im zweiten Akt. Entgegen bisheriger Interpretationen, in denen nur die Intrige des Prinzen und Marinelli dafür verantwortlich gemacht werden, dass Emilia um ihre Tugend fürchten und ihren Vater dazu bringen muss, sie umzubringen, wird dadurch die Abwertung allein durch die Avancen des Prinzen sichtbar, die später sowohl Emilias als auch für Odoardos Einschätzung der Ehrbarkeit Emilias in ihrem zukünftigen Leben bestimmen.

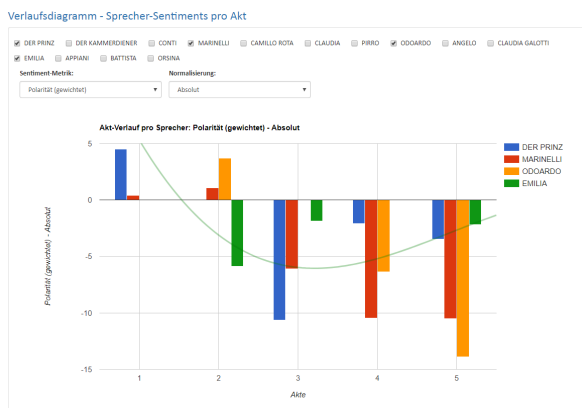


Abbildung 4: Polaritätsverlauf von Sprechern pro Akt – Emilia Galotti

Fazit

Insgesamt sind die ersten Analyse-Ergebnisse über das Web-Tool sehr vielversprechend. Dabei ist zu bedenken, dass über die Verwendung von SA-Lexika ein sehr einfacher SA-Ansatz gewählt wurde. Über ML- oder

Hybrid-Ansätze können Besonderheiten der poetischen und veralteten Sprache möglicherweise besser beachtet werden. Ferner ist fraglich, ob eine Reduktion auf das sonst in der SA übliche binäre System positiv/negativ ausreichend ist für komplexe Interpretationen von Emotionen in Dramen.

Durch Optimierung des SA-Verfahrens, Ausbau der Funktionen im Front-End und Erweiterung des Tools mit zusätzlichen Dramen sollen künftig Möglichkeiten und Nutzen der SA in der Dramenanalyse weiter exploriert werden.

Fußnoten

- <https://textgridrep.org/repository.html>; Hinweis: alle im Beitrag erwähnte URLs wurden zuletzt am 12.1.2018 überprüft
- Die vollständige Tabelle ist online verfügbar unter <https://drive.google.com/open?id=1cvyqiiLJ03XT1VNaWgSDoajeTE3wgeqxxr2PXp-VM4w>
- http://lauchblatt.github.io/QuantitativeDramenanalyseDH2015/FrontEnd/sa_selection.html

Bibliographie

Alm, Cecilia Ovesdotter / Sproat, Richard (2005): "Emotional sequencing and development in fairy tales.", in: *International Conference on Affective Computing and Intelligent Interaction* 668-674.

Alm, Cecilia Ovesdotter / Roth, Dan / Sproat, Richard (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* 579-586.

Buechel, Sven / Hellrich, Johannes / Hahn, Udo (2017): "The Course of Emotion in Three Centuries of German Text – A Methodological Framework.", in: *Digital Humanities 2017* 176-179.

Clematide, Simon / Klenner, Manfred (2010): "Evaluation and extension of a polarity lexicon for German.", in: *Proceedings of the First Workshop on Computational Approaches to Subjectivity and Sentiment Analysis* 7-13.

D'Andrea, Alessia et al. (2015): "Approaches, tools and applications for sentiment analysis implementation.", in *International Journal of Computer Applications* 125.3: 26-33.

De Smedt, Tom / Daelemans, Walter (2012): "Pattern for python.", in: *Journal of Machine Learning Research* 13: 2063-2067.

Eger, Steffen / Gleim, Rüdiger / Mehler, Alexander. (2016). "Lemmatization and Morphological Tagging in German and Latin: A Comparison and a Survey of the State-of-the-art.", in: *LREC* 1507-1513.

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

Fleiss, Joseph L. (1971): "Measuring nominal scale agreement among many raters.", in: *Psychological bulletin* 76.5: 378-382.

Gonçalves, Pollyanna, et al. (2013): "Comparing and combining sentiment analysis methods.", in: *Proceedings of the first ACM conference on Online social networks* 27-33.

Jannidis, Fotis, et al. (2016): "Analyzing Features for the Detection of Happy Endings in German Novels.", in: *arXiv preprint arXiv:1611.09028*

Jurish, Bryan (2011): *Finite-state canonicalization techniques for historical German*. Diss. Universitätsbibliothek der Universität Potsdam.

Kakkonen, Tuomo / Kakkonen, Gordana Gali# (2011): "SentiProfiler: creating comparable visual profiles of sentimental content in texts.", in: *Language Technologies for Digital Humanities and Cultural Heritage* 62-67.

Kennedy, Alistair / Inkpen, Diana (2006): "Sentiment classification of movie reviews using contextual valence shifters.", in: *Computational intelligence* 22.2: 110-125.

Kim, Evgeny / Padó, Sebastian / Klinger, Roman (2017): "Investigating the relationship between Literary Genres and Emotional Plot Development.", in: *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* 17-26.

Kouloumpis, Efthymios / Wilson, Theresa / Moore, Johanna D. (2011): "Twitter sentiment analysis: The good the bad and the omg!.", in: *Proceedings of the Fifth International Conference on Weblogs and Social Media* 538-54.

Liu, Bing (2016): *Sentiment analysis: Mining opinions, sentiments, and emotions*. New York: Cambridge University Press.

McGlohon, Mary / Glance, Natalie S. / Reiter, Zach (2010) "Star Quality: Aggregating Reviews to Rank Products and Merchants.", in: *Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2010)* 114-121.

Mohammad, Saif (2011): "From once upon a time to happily ever after: Tracking emotions in novels and fairy tales.", in: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* 105-114.

Mohammad, Saif M. / Turney, Peter 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* 26-34.

Nalisnick, Eric T. / Baird, Henry S. (2013): "Character-to-character sentiment analysis in shakespeare's plays.", in:

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics 479-483.

Remus, Robert / Quasthoff, Uwe / Gerhard, Heyer (2010): "SentiWS-A Publicly Available German-language Resource for Sentiment Analysis.", in: *LREC* 1168-1171.

Saif, Hassan, et al. (2014): "On stopwords, filtering and data sparsity for sentiment analysis of twitter.", in: *Proc. 9th Language Resources and Evaluation Conference (LREC)* 810-817.

Saße, Günter (1993): *Liebe und Ehe: oder, wie sich die Spontaneität des Herzens zu den Normen der Gesellschaft verhält. Lessings Minna von Barnhelm*. Tübingen: Niemeyer.

Schmid, Helmut (1995): "Improvements in part-of-speech tagging with an application to German.", in: *Proceedings of the acl sigdat-workshop*.

Vinodhini, G. / Chandrasekaran, R. M. (2012): "Sentiment analysis and opinion mining: a survey.", in: *International Journal of Advanced Research in Computer Science and Software Engineering* 2.6: 282-292.

Võ, Melissa LH, et al. (2009): "The Berlin affective word list reloaded (BAWL-R) ", in: *Behavior research methods* 41.2: 534-538.

Waltinger, Ulli (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)*.