

Methodenbeitrag: Sentimentanalyse

Marie Flüh  ¹

1. Universität Hamburg

forTEXT

Thema:	Sentimentanalyse	DOI:	10.48694/fortext.3797
Jahrgang:	1	Ausgabe:	7
Erscheinungsdatum:	2024-07-10	Erstveröffentlichung:	2019-09-16 auf fortext.net
Lizenz:			open  access

Allgemeiner Hinweis: Rot dargestellte *Begriffe* werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.

1. Definition

Die oft synonym verwendeten Begriffe Sentimentanalyse, Stimmungsanalyse oder -erkennung und Opinion Mining (vgl. **Opinion Mining**) bezeichnen einen Teilbereich des Fachgebiets „Information Retrieval“ (vgl. **Information Retrieval**) (u. a. Teilbereich der Computerlinguistik und Informatik), in dem die gezielte Suche nach Informationen aus großen Textmengen im Fokus steht. Wie der Begriff Sentiment (aus dem Franz. *le sentiment* für Gefühl, Stimmung) bereits verrät, handelt es sich bei der Sentimentanalyse um die automatisierte Analyse (vgl. **Text Mining**) von in Texten dargestellten menschlichen Gefühlen, Empfindungen und/oder Meinungen, die verbalisiert und dadurch nach außen getragen werden. In diversen Anwendungen geht es v. a. darum, durch eine Sentimentanalyse ein Stimmungsbild zu erzeugen und die im Text manifestierten Meinungen und Gefühle herauszufiltern.

Die Sentimentanalyse ist – oft als kommerzielle Dienstleistung – fester Bestandteil diverser Kommunikationsunternehmen und gehört zum Methodenkanon der Sozialwissenschaften. Die Anwendungen betreffen u. a. die automatisierte Extraktion der in sozialen Netzwerken vermittelten Gefühlslagen, die Überwachung von Internet-Foren sowie die Erkennung von negativen Äußerungen (Flaming) und deren Entfernung oder die Analyse von Kundenfeedback und Online Reviews zur Produktverbesserung (Ortner 2014; Liu 2015). Neben die analytische Funktion der Stimmungsanalyse tritt eine prognostische: Sentimentanalysen werden ebenfalls eingesetzt, um Hypothesen über mögliche zukünftige menschliche Reaktionen – z. B. auf ein Produkt oder zu einem bestimmten Thema – anstellen zu können (Liu 2015, 3). Als spezifische Methode der Textanalyse ist die Sentimentanalyse in ein interdisziplinäres Forschungsfeld eingebettet, und übernimmt Techniken aus den Bereichen Natural Language Processing (vgl. **NLP**) und Data- bzw. Web-Mining (Zhang und Liu 2014).

2. Anwendungsbeispiel

Literaturwissenschaftliche Fragestellungen, die sich durch eine Sentimentanalyse beantworten ließen, können auf der Makro- wie auch der Mikroebene von Texten ansetzen. Auf Korpusebene (vgl. **Korpus**) kann z. B. untersucht werden, ob Jean Pauls Werke in der literaturkritischen Berichterstattung des 18. Jahrhunderts eher positiv oder eher negativ bewertet wurden. Wenn Sie ein einzelnes Dokument untersuchen möchten, kann auf einer übergeordneten Dokumentenebene z. B. nach den Stimmungen gefragt werden, die den Handlungsverlauf in Arthur Schnitzlers *Der grüne Kakadu* begleiten. Es können aber auch noch kleinere Einheiten betrachtet werden, indem auf der Satzebene z. B. analysiert wird, ob es sich bei Karl Moor aus Friedrich Schillers Drama *Die Räuber* um eine Figur handelt, die mit eher positiven oder negativen Sentiments in Verbindung steht. Die prognostische Funktion der Sentimentanalyse kann bspw. auf die Vorhersage eines glücklichen bzw. unglücklichen Endes von Erzählungen (Zehe u. a. 2016) oder auf Genreklassifikationen abzielen (Kim, Padó und Klinger 2017).

3. Literaturwissenschaftliche Tradition

Die Sentimentanalyse steht in der Tradition psychoanalytischer, linguistischer, sozial- und kulturwissenschaftlicher sowie literaturwissenschaftlicher Forschungsansätze, die sich mit dem Zusammenhang zwischen Sprache und Emotionen beschäftigen. Die Auseinandersetzung mit Emotionen ist als wichtiger Bestandteil von der Rhetorik oder der Poetik seit Platon und Aristoteles eine der „[...] größten Konfliktzonen des Denkens und der Wissenschaft [...]“ (Von Koppenfels und Zumbusch 2016, 1) und blickt folglich auf eine weitaus längere Traditionslinie zurück als der rein literaturwissenschaftlich geprägte Diskurs. Klassische Emotionstheorien, die eine Verbindung zwischen Philosophie, Emotionen und Dichtung berücksichtigen, existieren seit dem 17. Jahrhundert (Landweer und Renz 2008, 5).

Literarische Texte stellen einen Phänomenbereich dar, der eine besonders intensive Symbiose von Sprache

und Emotionen aufweist. Sowohl in narrativen Texten selbst als auch in der literaturwissenschaftlichen Interpretation spielt die Analyse von Emotionen und Empfindungen deshalb eine Rolle. Die unterschiedlichen Ansätze gehen i. d. R. entweder vom Leser oder vom Text aus und verfolgen dementsprechend ein gemeinsames Forschungsinteresse, bei dem Fragen nach der emotionalen Wirkung von Literatur den Ausgangspunkt bilden (Mellmann 2016, 158). So existieren zahlreiche Einzelfallstudien, in denen die sprachliche Repräsentation von Emotionen in literarischen Texten und deren Prozessualität thematisiert werden. Gängiges Vorgehen ist hierbei die Untersuchung einer oder mehrerer Emotionen innerhalb eines bestimmten Werks oder – deutlich unterrepräsentiert – größer angelegte Studien, in denen große Textkorpora hinsichtlich der Emotionsmanifestationen im Text und deren Konzeptualisierungen untersucht werden (vgl. Hufnagel 2013 über Trauer in der Nibelungensage; Hindinger 2013 über Leid und Qual in der Literatur des 18. und 19. Jahrhunderts; Wieland 2013 über Emotionen in Max Frischs *Homo Faber* und *Montauk*; größer angelegte Studien: Alfes (1995) über Gefühle als Bestandteile literarischer Prozesse oder Winko (2003) über Emotionen in Gedichten und nicht zuletzt das Handbuch *Literatur und Emotionen*, in dem unterschiedliche theoretische Zugänge zum Zusammenspiel beider Bereiche erläutert werden).

Literaturwissenschaftliche Vorläufer der Sentimentanalyse finden sich auch in der Rezeptionsästhetik. Die Frage danach, wie künstlerische Werke wahrgenommen werden, ob die Möglichkeit einer emotionalen Wahrnehmung im Werk selbst angelegt ist oder sich erst im Prozess der Rezeption entfaltet (Winko 2003), setzt u. a. eine Untersuchung des emotiven bzw. expressiven oder emotionalen Gehalts literarischer Texte voraus. Die drei – leider oft synonym verwendeten Begriffe – beziehen sich auf unterschiedliche Ebenen, auf denen Emotionen analysiert werden: Varianten des Begriff *emotional* sind v. a. der Beschreibung und Analyse psychischer Phänomene vorbehalten, der Terminus *emotiv* bezeichnet Kategorien des Emotionalen in der Sprache und *expressiv* betrifft die Intensität eines emotionalen Gefühlsausdrucks (Ortner 2014, 68).

Im Kern lassen sich drei Phasen der literaturwissenschaftlichen Behandlung von Emotionen unterscheiden. Bis in die 1960er Jahre wurde die Forschung von der „Autorität des einfühlenden Interpreten“ (Winko 2003, 10) geprägt, der aufgrund seiner Expertenkompetenz Aussagen über einen Text und die durch ihn bekundeten Gefühle tätigte. Im Zuge der szenischen Wende in den 70er Jahren wurden Emotionen als subjektives und deshalb nicht objektivierbares Phänomen aufgefasst, das sich nicht literaturwissenschaftlich erforschen lässt. Der Zusammenhang zwischen Gefühlen, Emotionen und Literatur wurde aufgrund dieser vermeintlich charakteristischen Unzugänglichkeit in der Literaturwissenschaft der 70er und 80er Jahre kaum thematisiert. Weder in strukturalistisch fundierten Analysen literarischer Texte, die v. a. den kognitiven Gehalt von Literatur behandelten, noch in philosophischen oder wissenschaftlichen Bezugstheorien wurde das Gefühlsthema aufgegriffen. Ab den 90er Jahren wurden Emotionen als Thema literarischer und nicht-literarischer Texte, aber auch in der Psychologie, der Philosophie und der Soziologie, wieder häufiger untersucht (Winko 2003, 10f.).

Die hieran anschließende Forschung zum Verhältnis zwischen Literatur und in textueller Gestalt erscheinenden Emotionen lässt sich in produktions- oder rezeptionsbezogen argumentierende Forschungsarbeiten gruppieren. Es finden sich u. a. empirische literaturwissenschaftliche Forschungsansätze (Alfes 1995), poststrukturalistische Arbeiten (Keitel 1996; Barthes 1973), formalistische bzw. strukturalistische Ansätze (Tomaševskij 1985) und kontextbezogene Ansätze (Fick 1993; Meier 1993), die meistens auf eine Gesamtdeutung oder die Deutung einzelner Textstellen abzielen. Einen systematischen Forschungsüberblick sämtlicher literaturwissenschaftlicher Ansätze, die sich mit dem Thema Literatur und Emotionen auseinandersetzen (produktionsbezogene und rezeptionsbezogene Ansätze, imaginationstheoretische Modelle, Affektlehre, Rhetorik, die Kitsch-Debatte, die philosophische Debatte im angelsächsischen Raum, poststrukturalistische Ansätze, formalistische, strukturalistische und poststrukturalistische Ansätze) liefert Winko (2003). Textbezogene Ansätze behandeln Emotionen als explizite (also z. B. durch eine Figur oder die Erzählinstanz konkret thematisierte) oder implizite (also z. B. durch eine Handlung oder das Verhalten einer Figur indirekt ausgedrückte) Größe. Obwohl gerade in literarischen Texten Emotionen v. a. implizit vorliegen, dominieren explizite Ansätze (Winko 2003, 47).

4. Diskussion

Die digitale Literaturwissenschaft knüpft zum einen an diese Traditionslinien an und greift außerdem – als ein per se interdisziplinär ausgerichteter Fachbereich – auf die methodische und konzeptuelle Arbeit aus Informatik und Informationswissenschaft zurück. Beispiele für auf literarischen Texten basierende Sentimentanalysen stellen die in Abbildung 1 und 2 dargestellten Untersuchungen literarischer Texte hinsichtlich des Zusammenhangs zwischen Handlungsverlauf und Emotionen (Jockers 2014; Jockers 2015) dar.

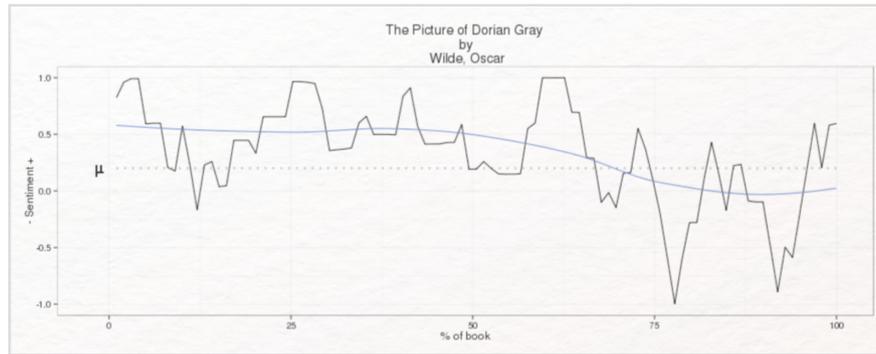


Abb. 1: Emotionsverlauf in Oscar Wildes' Roman The Picture of Dorian Gray (Jockers 2015)

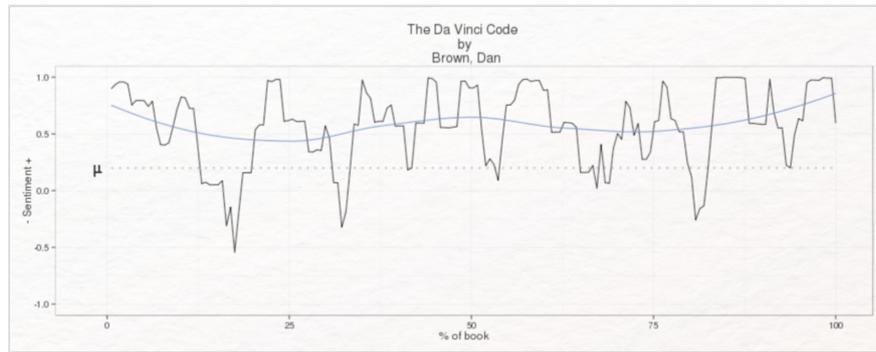


Abb. 2: Das Sentiment-Profil eines page turners: Dan Browns Roman The Da Vinci Code (Jockers 2015)

Mit der Bestimmung des Sentimentgehalts ausgewählter Märchen (Mohammad 2011), der Ermittlung prototypisch verlaufender Stimmungsbögen in fiktionalen Texten aus dem Project Gutenberg (Reagan u. a. 2016) und Sentimentanalysen zu populären Romanen wie *Harry Potter* oder zu Erzählungen Franz Kafkas (siehe Abb. 3) liegen weitere Sentimentanalyse mit literaturwissenschaftlichem Schwerpunkt vor.

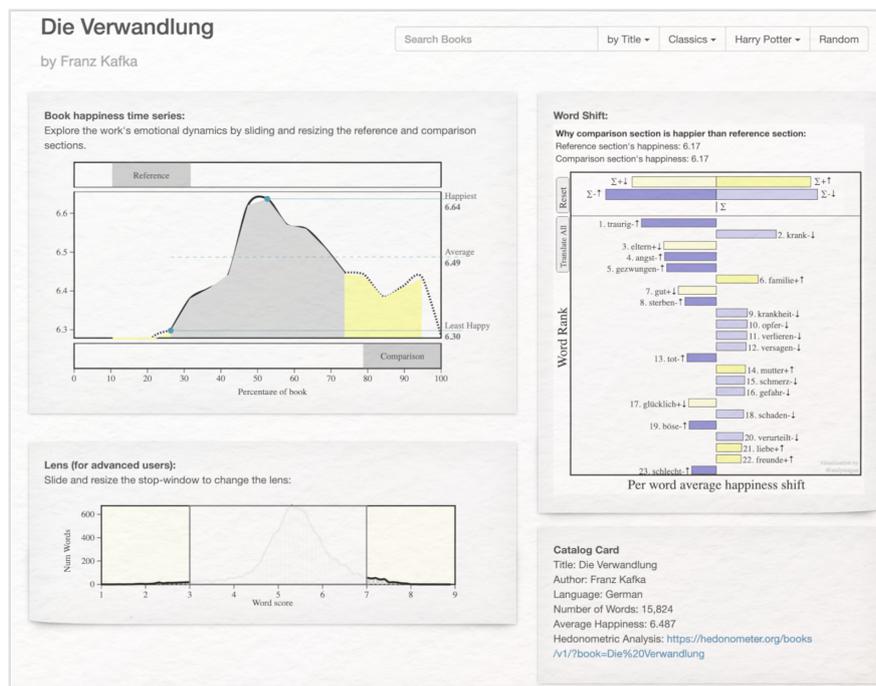


Abb. 3: Sentimentanalyse zu Franz Kafkas Die Verwandlung, bereitgestellt vom Computational Story Lab (https://hedonometer.org/timeseries/en_all/)

Ein weiteres anschauliches Beispiel für die quantitative Analyse (vgl. **Distant Reading**) ausgewählter Dramen Gotthold Ephraim Lessings – mit stärkerer Ausrichtung auf die technische Entwicklung – hinsichtlich der hier enthaltenen Emotionen (Schmidt 2017) ist in *Katharsis* (Flüh 2024a) implementiert (siehe Abb. 4). Für einen Überblick über den Forschungsstand zur Sentiment- und Emotionsanalyse zur Analyse von Literatur: vgl. Kim und Klinger (2019).

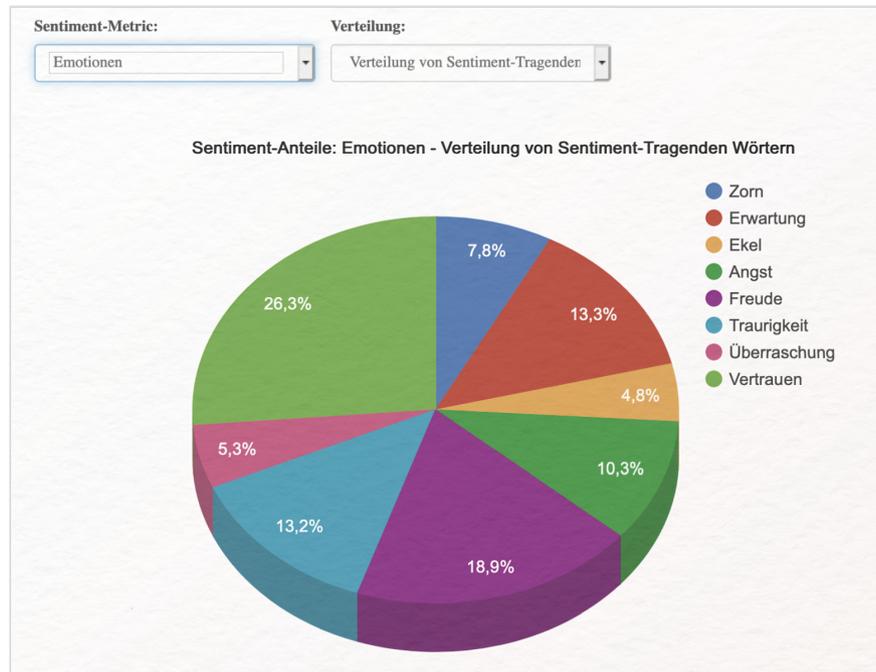


Abb. 4: Über *Katharsis* bezogenes Kreisdiagramm zu in Lessings Drama *Miss Sara Sampson* enthaltenen Basisemotionen

Während es recht unterschiedliche Antworten – z. B. aus philosophischer, neurologischer, psychologischer oder eben aus literaturwissenschaftlicher Perspektive – auf die Frage gibt, wie Gefühle in den Text kommen (Winko 2003), sind die Antwortmöglichkeiten auf die Frage, auf welche Weise Gefühle computerbasiert wieder aus Texten herausgefiltert werden können, relativ überschaubar. Um Sentimentanalysen durchzuführen, stehen zwei unterschiedliche Ansätze zur Verfügung, die z. T. auch miteinander kombiniert werden: lexikon-basierte Sentimentanalysen und Sentimentanalysen, bei denen Methoden des maschinellen Lernens (vgl. **Machine Learning**) zum Einsatz kommen.

Die Lexika wurden i. d. R. manuell erstellt und knüpfen an unterschiedliche Konzepte zur Messung von Emotionen an (Lehmann 2017). In den meisten Fällen werden die Lexika als Datensätze zur freien Verfügung gestellt. Eine laufende Bibliografie der für die Sentimentanalyse geeigneten Lexika fehlt leider. Eine hilfreiche Matrix der besonders umfangreichen bzw. etablierten Lexika bieten z. B. Lehmann (2017) oder Schmidt (2017). Die Vielzahl an unterschiedlichen Sentimentwörterbüchern (vgl. **Sentimentwörterbuch**; siehe Ressourcen am Ende dieses Methodeneintrags) verweisen auf die Tatsache, dass sich Emotionen, Stimmungen und Wertungen je nach Textsorte, -sprache und -kontext unterscheiden. Sentimentwörterbücher sind domänenspezifische Konstrukte. Beispiele aus Goethes *Faust I* verdeutlichen die Nachteile der lexikonbasierten Sentimentanalyse. Orthographische Besonderheiten führen dazu, dass einige Wörter in einem automatisierten Abgleich mit einem Sentimentwörterbuch nicht erkannt werden. Das Wort „Bewundrung“ würde bspw. aufgrund der von Goethe verwendeten Orthographie nicht als emotionsgeladenes Wort erkannt. Mit Schwierigkeiten verbunden ist außerdem ein spezifischer Wortschatz aus Wörtern, die im zeithistorischen Kontext des 18. und 19. Jahrhunderts verwendet wurden, in den Sentimentwörterbüchern – die zum größten Teil auf ein zeitgenössisches Vokabular abgestimmt sind – aber nicht vorkommen. Das führt bei Sentimentanalysen sozialer Medien zu Erkennungsraten zwischen 80 und 95 %. Da literarische Texte aber gerade ein historischer Sprachstil oder eine poetische Sprache auszeichnen, kann die Ausgrenzung dieser Eigenschaften zu Fehlkalkulationen führen (Schmidt, Burghardt und Dennerlein 2018a, 4f.). „bebaglich“, „Gönner“, „holden“, „Dirne“, „Lebensglück“ oder „ergetzen“ sind bspw. in zeitgenössischen Sentimentwörterbüchern wie SentiWS nicht enthalten.

Darüber hinaus handelt es sich bei der Wortsemantik nicht um eine statische Größe, sondern um eine variable Einheit, die sich über die Zeit verändern kann. „Dirne“, das im *Faust* ein junges, anmutiges und schönes Mädchen bezeichnet, ist in diesem Kontext positiv konnotiert. Im modernen deutschen Sprachgebrauch kann „Dirne“ allerdings auch eine Prostituierte bezeichnen und ist eher negativ konnotiert. In Anbetracht der Schwierigkeit, das historische Vokabular mittels lexikonbasierter Sentimentanalyse zu berücksichtigen, stellt sich die Frage,

wie sehr das Auslassen dieser Wörter das Gesamtergebnis der Analyse verfälschen würde. Erste Beispielanalysen zeigen eine konsistente Verbesserung durch Lexikonerweiterung mit historisch-linguistischen Varianten. Auf diese Weise wird versucht, die Zuverlässigkeit bei der lexikonbasierten Erkennung des historischen Vokabulars zu erhöhen. Eine verbesserte Lemmatisierung (vgl. **Lemmatisieren**) wirkt sich positiv auf die Leistung aus (Schmidt, Burghardt und Dennerlein 2018a).

Ein weiteres Problem der lexikonbasierten Sentimentanalyse stellen Negationen dar, deren Auftreten das Analyseergebnis verfälschen kann. Die Implementierung von Tools, die Negationen erkennen können, würde zu besseren Ergebnissen führen, verlangt aber weitere technische Vorkenntnisse. Darüber hinaus wird deutlich, dass es spezifischer Sentimentwörterbücher bedarf, welche historische und orthographische Besonderheiten genauso einbeziehen wie individuelle Schreibstile. Für die literaturwissenschaftlich orientierte Sentimentanalyse gibt es bisher nur vereinzelte Ansätze, die auf die Erstellung eigener textsortenspezifischer Lexika abzielen. Lexikonbasierte Ansätze erfassen lediglich emotionsausdrückende Lexeme und deren **Sentimentwert**. Die Kluft zwischen Alltag und Kunstwerk bzw. einer außerkünstlerischen und alltäglichen Sprache auf der einen und einer innerkünstlerischen auf der anderen Seite (Herding 2008, 10) verweist auf ein Problemfeld der Methode, das sich dann ergibt, wenn versucht wird, zwei sehr unterschiedliche Untersuchungsgegenstände mit identischen Untersuchungsverfahren zu explorieren. Eine kontextuelle Emotionsthematisierung oder die Erkennung von emotionsausdrückenden Metaphern und expressiven Sprechakten lassen sich auf diese Weise nicht erkennen. Auch andere sprachliche Besonderheiten wie Ironie, Sarkasmus oder mehrdeutige Wörter werden in den lexikonbasierten Ansätzen nicht berücksichtigt.

Um Meinungen, Gefühle und Emotionen von Menschen zu analysieren oder vorauszusagen, können außerdem Verfahren des überwachten maschinellen Lernens eingesetzt werden. Diese Methode basiert auf einem mit Sentiment-Informationen angereicherten Trainingskorpus. Für Texteinheiten werden die Sentimente – i. d. R. per Hand – annotiert (vgl. **Annotation**; Schmidt, Burghardt und Dennerlein (2018a); Ignatow und Mihalcea (2017), 150). Anders als im Fall der lexikonbasierten Analyseverfahren wird durch die manuelle Annotation (Jacke 2024) also ein Modell trainiert, welches die Besonderheiten der Darstellung von Emotionen in literarischen Texten berücksichtigen kann. Zum jetzigen Stand der Forschung stellen sich Ansätze aus dem Bereich des maschinellen Lernens als leistungsstärker heraus (Schmidt, Burghardt und Dennerlein 2018a). Dennoch sind Beiträge der zweiten „Großform“ der Sentimentanalyse in der Minderheit, da es bis jetzt an der wichtigsten Grundlage hierfür mangelt: mit Sentiment-Informationen angereicherte Trainingskorpora. Bei einem Großteil der literaturwissenschaftlichen Sentimentanalysen wird deshalb auf lexikonbasierte Ansätze zurückgegriffen, obwohl sich neben den genannten Problemen – zumindest für den deutschen Sprachraum – ein weiterer Mangel abzeichnet: Es existieren deutlich weniger evaluierte Ressourcen für die lexikonbasierte Sentimentanalyse als etwa im angelsächsischen Sprachraum.

Für beide angeführten Formen der Sentimentanalyse gilt: Der Phänomenbezeichnung und der Frage danach, was genau mit der Sentimentanalyse überhaupt „gemessen“ wird, fehlt es oft an terminologischer Genauigkeit. Während Jockers (2015) von „emotional content“, „sentiment“ oder „emotional valence“ spricht, ist an anderer Stelle von „subjectivity attitude“ (Zehe u. a. 2016), „emotion words“ und deren Dichte (Mohammad 2011) oder negativen und positiven „emotionalen Bewertungen“ (Schmidt, Burghardt und Dennerlein 2018a) die Rede. Oft bleibt es zunächst relativ undeutlich, welches Phänomen – Gefühl, Affekt, Empfindung, Emotion, Leidenschaft, Eindruck oder Stimmung – mit der Sentimentanalyse literarischer Texte untersucht werden soll. Es fehlt eine geeignete Metasprache, deren konsistente Verwendung terminologische und sachliche Verwirrungen vermeiden würde und die den metatheoretischen Rahmen nutzen – wie ihn etwa Simone Winko für die literaturwissenschaftliche Emotionsforschung oder Monika Schwarz-Friesel für die linguistische Emotionsforschung – bereits eingeführt haben.

Eine in die Tiefe gehende Forschung müsste außerdem Ansätze zur Erklärung emotionaler Phänomene als theoretisches Fundament einbeziehen und eine Kontextualisierung vornehmen, um die Ergebnisse der Sentimentanalyse nicht im „luftleeren“ Raum stehen zu lassen. Dieser Rückbezug fordert Forschende außerdem heraus, die literaturwissenschaftliche, inhaltliche Relevanz von Sentimentanalysen in den Diskurs zu holen. Allerdings wird der Blick auf die inhaltliche Relevanz und die Entwicklung einer Metasprache durch die Fokussierung auf eine technische Weiterentwicklung und Evaluation bestehender Methoden beider Großformen der Sentimentanalyse zum Teil verstellt. Gleichzeitig ist die Grundlagenforschung ein notwendiger Schritt in Richtung einer literaturwissenschaftlichen Domänenadaption (vgl. **Domänenadaption**) der Sentimentanalyse. Angesichts des interdisziplinär verwurzelten Untersuchungsgegenstands – Emotionen und Gefühle – liegt es nahe bzw. ist es gerade in den interdisziplinär ausgerichteten digitalen Geisteswissenschaften gängige Praxis, methodisches Vorgehen aus verwandten Fachdisziplinen auf die Auseinandersetzung mit „unserem Material“ zu übertragen. Die literaturwissenschaftlich ausgerichtete Sentimentanalyse stößt hierbei allerdings an deutliche Grenzen, da Untersuchungsgegenstand und Untersuchungsverfahren (noch) nicht kompatibel erscheinen.

5. Technische Grundlagen

Tools zur Sentimentanalyse, die über eine grafische Benutzeroberfläche (vgl. **GUI**) verfügen, ohne ausgeprägtes technisches Vorwissen bedienbar sind, eine individuelle Textauswahl zulassen und der Spezifik literarischer

Texte gerecht werden, sind noch nicht entwickelt worden. Gleichzeitig müssten verstärkt Anforderungsanalysen durchgeführt werden, die den Einbezug literaturwissenschaftlicher Expertise in die technische Entwicklung ermöglichen. Ohne die Entwicklung einer grafischen Benutzeroberfläche wird die literaturwissenschaftliche Sentimentanalyse eine Form der Textanalyse bleiben, die erhebliches technisches Vorwissen verlangt. Das gilt sowohl für die Anwendung lexikonbasierter Verfahren als auch für Verfahren, die maschinelles Lernen beinhalten. Mit Tools wie Weka oder RapidMiner stehen zwar Machine-Learning-Tools mit einem GUI zur Verfügung. Da es sich hier allerdings um Werkzeuge zum Machine-Learning handelt, kann von einem niedrigschwelligen Zugang nicht die Rede sein.

Elementare Bestandteile der lexikon-basierten Sentimentanalyse sind annotierte Satz- und/oder Wortlisten (sog. Sentimentwörterbücher) und eine Aggregationsmethode (z. B. hierfür entworfenen Python- oder R-Programme). Die Sentimentwörterbücher oder Wortlisten enthalten Wörter, die mit Informationen über deren Emotionsgehalt angereichert wurden (*emotion label*). Gängig ist z. B. die Angabe der Polarität eines Wortes, also deren Auszeichnung als negativ, neutral oder positiv, in einem bestimmten Intervall. Die lexikonbasierte Sentimentanalyse wird in der Literaturwissenschaft häufiger eingesetzt. Hierbei werden sämtliche in einem Text vorkommende Wörter mit der jeweiligen Wortliste oder einer Kombination aus mehreren Sentimentwörterbüchern abgeglichen. Sentimentwörter (*Sentiment Bearing Word*, SBW) werden als solche erkannt und bekommen durch den Abgleich mit dem Lexikon einen Sentimentwert zugewiesen. Eine frei verfügbare und im deutschsprachigen Raum besonders häufig verwendete bzw. etablierte Ressource für die lexikonbasierte Sentimentanalyse ist der SentimentWortschatz (kurz: SentiWS) der Universität Leipzig. Für die hier enthaltenen Wörter – die *Sentiment Bearing Words* – werden jeweils eine positive oder negative Polarität im Intervall [-1; 1], die Wortart und mögliche Flexionsarten angegeben. Die im Jahr 2018 zuletzt aktualisierte Version enthält insgesamt circa 18.000 negative und 16.000 positive Wortformen; circa 1.800 negative und 1.650 positive Grundformen und unterschiedliche Flexionsformen. Neben Adverbien und Adjektiven sind auch Nomen und Verben, die die Träger von Sentiment sind, verzeichnet (Remus, Quasthoff und Heyer 2010). Laden Sie SentiWS herunter, so erhalten sie zwei UTF-8 (vgl. Unicode/UTF-8) kodierte TXT-Dateien (vgl. Reintext-Version), in denen jeweils die positiven und negativen SBWs in alphabetischer Reihenfolge aufgelistet sind (siehe Abb. 5).

Ambivalenz NN	-0.3266	Ambivalenzen
Androhung NN	-0.0048	Androhungen
Anfälligkeit NN	-0.0048	Anfälligkeiten
Angreifer NN	-0.0048	Angreifers, Angreifer
Angriff NN	-0.212	Angriffe, Angriffen, Angriffs, Angriffes
Angst NN	-0.514	Ängste, Ängsten
Anklage NN	-0.3468	Anklagen
Anmaßung NN	-0.0048	Anmaßungen
Annullierung NN	-0.0048	Annullierungen
Anomalie NN	-0.0048	Anomalien
Anspannung NN	-0.0048	Anspannungen
Anstrengung NN	-0.0048	Anstrengungen
Antipathie NN	-0.0048	Antipathien
Arbeitslose NN	-0.3549	Arbeitslosen, Arbeitsloser, Arbeitslose
Arbeitslosigkeit NN	-0.3654	
Armut NN	-0.2038	
Aufstockung NN	0.0040	Aufstockungen
Auftrag NN	0.0040	Aufträgen, Aufträge, Auftrages, Auftrags, Aufträge
Auftrieb NN	0.0040	Auftrieben, Auftriebs, Auftriebes, Auftriebe
Aufwertung NN	0.0040	Aufwertungen
Aufwärtstrend NN	0.0040	Aufwärtstrends
Augenweide NN	0.0844	Augenweiden
Ausbau NN	0.0040	Ausbaues, Ausbaus
Ausbildung NN	0.0040	Ausbildungen
Ausdauer NN	0.0886	
Ausgleich NN	0.0040	Ausgleichen, Ausgleichs, Ausgleiches, Ausgleiches
Ausgleichszahlung NN	0.0040	Ausgleichszahlungen
Auszeichnung NN	0.0040	Auszeichnungen
Authentizität NN	0.0040	
Autonomie NN	0.0040	Autonomen
Bedeutung NN	0.0040	Bedeutungen
Befreiung NN	0.0040	Befreiungen

Abb. 5: Ausschnitt der Liste negativer und positiver Sentimentwörter: Wort, POS-Tag, Ausprägung und Flexionsformen (-1: sehr negativer, +1: sehr positiver Sentiment Wert)

Ein POS-Tagging (vgl. POS) wurde auf Basis des Stuttgart-Tübingen-Tagsets (STTS) durchgeführt. Ausgezeichnet wurden Nomen (NN), attributive und deskriptive Adjektive (ADJX), infinite Verben (VVINF) und Adverbien (ADV). Die Flexionsformen wurden aus einer projektinternen Datenbank bezogen.

		Positive	Negative
Adjectives	Baseforms	784	698
	Inflections	11,782	10,604
Adverbs	Baseforms	6	4
	Inflections	0 ³	0 ³
Nouns	Baseforms	584	686
	Inflections	521	806
Verbs	Baseforms	312	430
	Inflections	2,453	3,100
All	Baseforms	1,650	1,818
	Inflections	14,756	14,510
	Total	16,406	16,328

Abb. 6: Zusammensetzung des SentimentWortschatz (SentiWS)

Dem Wörterbuch liegen die Auswertung und Zusammenführung dreier Quellen zugrunde. Die Ausprägung der einzelnen Wörter wurden mittels der „Pointwise Mutual Information“-Methode (PMI) berechnet. Bei diesem Ansatz wird die semantische Ausprägung eines Wortes (SO: *semantic orientation*) aus semantischen Assoziationen des Wortes (SA: *semantic association*) zu ausgewählten Wörtern (sog. *seed words*: ausgewählte Wörter, die bei der Berechnung quasi als exemplarische Schlüsselwörter verwendet werden), die stark positiv oder negativ konnotiert sind, abgeleitet. Die semantische Ausprägung eines Wortes bezeichnet folglich den Grad der Verbindung eben dieses Wortes zu einem manuell erstellten Set an Wörtern, die positiv oder negativ konnotiert sind. Für jedes Wort wird berechnet, ob und wie stark es mit positiven und negativen Wörtern assoziiert wird. (Die für die Berechnung der SO eines Wortes herangezogenen positiven *seed words* sind „gut“, „schön“, „richtig“, „glücklich“, „erstklassig“, „positiv“, „großartig“, „ausgezeichnet“, „lieb“, „exzellent“ und „phantastisch“. Die für die Berechnung der SO eines Wortes herangezogenen negativen *seed words* sind „schlecht“, „unschön“, „falsch“, „unglücklich“, „zweitklassig“, „negativ“, „scheiße“, „minderwertig“, „böse“, „armselig“ und „mies“). Die in Abbildung 5 und 6 angegebenen Kommazahlen stellen das Ergebnis dieses Berechnungsverfahrens dar und geben für jedes Wort einen Sentimentwert an. Hierbei ist +1 der positivste Sentimentwert und -1 der negativste Sentimentwert, der einem Wort zugeordnet werden kann. Sehr positive Wörter sind beispielsweise „Freude“, mit einem Sentimentwert von 0,6502 oder „zuvorkommend“ mit einem Sentimentwert von 0,6669. Mit Sentimentwerten von -0,9269 und -0,4889 („schädlich“ bzw. „Aggression“) handelt es sich um besonders negative Sentimentwörter.

Externe und weiterführende Links

- Hedonometer: https://web.archive.org/save/https://hedonometer.org/timeseries/en_all/ (Letzter Zugriff: 04.09.2024)
- Weka: <https://web.archive.org/save/https://www.cs.waikato.ac.nz/ml/index.html> (Letzter Zugriff: 04.09.2024)
- RapidMiner: <https://web.archive.org/save/https://rapidminer.com/solutions/text-mining/> (Letzter Zugriff: 04.09.2024)
- Python: <https://web.archive.org/save/https://www.python.org> (Letzter Zugriff: 04.09.2024)
- R-Project: <https://web.archive.org/save/https://www.r-project.org/> (Letzter Zugriff: 04.09.2024)
- SentimentWortschatz: <https://web.archive.org/save/http://wortschatz.uni-leipzig.de/de/download> (Letzter Zugriff: 04.09.2024)

Bibliographie

- Alfes, Henrike F. 1995. *Literatur und Gefühl: Emotionale Aspekte literarischen Schreibens und Lesens*. Opladen: Westdeutscher Verlag.
- Barthes, Roland. 1973. *Le plaisir du texte*. Paris: Éd. du Seuil.
- Böschstein, Hermann. 1954. *Deutsche Gefühlskultur: Studien zu ihrer dichterischen Gestaltung*. Bern: Haupt.

- Clausen, Rosemarie. 1960. *Gustaf Gründgens Faust in Bildern. Nach einer Verfilmung der Goetheschen Tragödie*. Braunschweig: Westermann.
- Fick, Monika. 1993. *Sinnenwelt und Weltseele: Der psychophysische Monismus in der Literatur der Jahrhundertwende*. Tübingen: Niemeyer.
- Flüh, Marie. 2024a. Ressourcenbeitrag: Katharsis. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 7. Sentimentanalyse (7. Oktober). doi: 10.48694/fortext.3795, <https://fortext.net/ressourcen/textsammlungen/katharsis>.
- . 2024b. Toolbeitrag: LIWC. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 7. Sentimentanalyse (7. Oktober). doi: 10.48694/fortext.3800, <https://fortext.net/tools/tools/liwc>.
- . 2024c. Toolbeitrag: SentText. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 7. Sentimentanalyse (7. Oktober). doi: 10.48694/fortext.3799, <https://fortext.net/tools/tools/senttext>.
- Herding, Klaus. 2008. Wie sich Gefühle Ausdruck verschaffen. Emotionen in Nahsicht. In: *Wie sich Gefühle Ausdruck verschaffen. Emotionen in Nahsicht*, hg. von Krause Herding und Antje Krause-Wahl, 7–16. Berlin: Driesen.
- Ignatow, Gabriel und Rada Mihalcea. 2017. *Text Mining. A guidebook for the Social Sciences*. Los Angeles (u.a.): SAGE.
- . 2018. *An Introduction to Text Mining. Research Design, Data Collection, and Analysis*. Los Angeles (u.a.): SAGE.
- Jacke, Janina. 2024. Methodenbeitrag: Manuelle Annotation. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3748, <https://fortext.net/routinen/methoden/manuelle-annotation>.
- Jockers, Matthew. 2014. A Novel Method for Detecting plot. *Matthew L. Jockers*. <http://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/> (zugegriffen: 16. Juli 2019).
- . 2015. Revealing Sentiment and Plot Arcs with the Syuzhet Package. *Matthew L. Jockers*. <http://www.matthewjockers.net/2015/02/02/syuzhet/> (zugegriffen: 16. Juli 2019).
- Keitel, Evelyn. 1996. *Von den Gefühlen beim Lesen: Zur Lektüre amerikanischer Gegenwartsliteratur*. München: Fink.
- Kim, Evgeny und Roman Klinger. 2019. A Survey on Sentiment and Emotion Analysis for Computational Literary Studies. *Zeitschrift für digitale Geisteswissenschaften*. doi: 10.17175/2019_008, http://www.zfdg.de/2019_008 (zugegriffen: 2. Mai 2020).
- Kim, Evgeny, Sebastian Padó und Roman Klinger. 2017. Prototypical Emotion Development in Literary genres. In: *Proceedings of the Joint SIGHUM Workshop on Computational Linguistic for Cultural Heritage, Social Sciences, Humanities and Literature*, 17–26.
- Kiritchenko, Svetlana, Xiaodan Zhu und Saif M. Mohammad. 2014. Sentiment Analysis in Short Informal Texts. *Journal of Artificial Intelligence Research* 50: 723–762. doi: 10.1613/jair.4272,.
- Landweer, Hilge und Ursula Renz. 2008. Zur Geschichte philosophischer Emotionstheorien. In: *Handbuch klassische Emotionstheorien*, hg. von Hilge Landweer und Ursula Renz, 1–18. Berlin, New York: de Gruyter.
- Lehmann, Jörg. 2017. Quantifizierung von Emotionswörtern in Texten: 34.
- Liu, Bing. 2015. *Sentiment Analysis: Mining Opinions, Sentiments and Emotions*. Cambridge: University Press.
- Meier, Albert. 1993. *Die Dramaturgie der Bewunderung: Untersuchungen zur politisch-klassizistischen Tragödie des 18. Jahrhunderts*. Frankfurt am Main: Klostermann.
- Mellmann, Katja. 2016. Empirische Emotionsforschung. In: *Handbuch Literatur & Emotionen*, hg. von Martin von Koppenfels und Cornelia Zumbusch, 158–175. Berlin, Boston: de Gruyter.
- Mohammad, Saif M. 2011. NRC Word-Emotion Association Lexicon (aka EmoLex). <https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm> (zugegriffen: 1. September 2019).
- Ortner, Heike. 2014. *Text und Emotionen: Theorie, Methode und Anwendungsbeispiele emotionslinguistischer Textanalyse*. Tübingen: Narr.
- Reagan, Andrew J, Lewis Mitchell, Dilan Kiley, Christopher M Danforth und Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science* 5, Nr. 1 (Dezember): 31. doi: 10.1140/epjds/s13688-016-0093-1, <http://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0093-1> (zugegriffen: 10. Dezember 2019).
- Remus, R., U. Quasthoff und G. Heyer. 2010. SentiWS - A Publicly Available German-language Resource for Sentiment Analysis. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 1168–1171. Valetta, Malta: European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2010/pdf/490_Paper.pdf.
- Schmidt, Thomas. 2017. 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. Regensburg: Universität Regensburg.
- Schmidt, Thomas, Manuel Burghardt und Katrin Dennerlein. 2018b. „Kann man denn auch nicht lachend sehr ernsthaft sein?“ - Zum Einsatz von Sentiment Analyse-Verfahren für die quantitative Untersuchung von Lessings Dramen. In: *Book of Abstracts, DHd 2018*. https://epub.uni-regensburg.de/37579/1/Self-Archiving-Version_DHd-2018.pdf.
- . 2018a. „Kann man denn auch nicht lachend sehr ernsthaft sein?“-Zum Einsatz von sentiment analyse-verfahren für die quantitative untersuchung von lessings dramen. In: *Book of Abstracts, DHd 2018*. Köln. https://epub.uni-regensburg.de/37579/1/Self-Archiving-Version_DHd-2018.pdf (zugegriffen: 9. September 2019).
- Schmidt, Thomas, Manuel Burghardt und Christian Wolff. 2018. Herausforderungen für Sentiment Analysis bei

- literarischen Texten. In: *INF-DH 2018*, hg. von Manuel Burghardt und Claudia Müller-Birn, Workshopband: Bonn: Gesellschaft für Informatik e.V. doi: 10.18420/inf2018-16,.
- Tomaševskij, Boris. 1985. *Theorie der Literatur*. Wiesbaden: Harrassowitz.
- Von Koppenfels, Martin und Cornelia Zumbusch. 2016. Einleitung. Literatur und Emotionen. In: *Handbuch Literatur & Emotionen*, hg. von Hilge Landweer und Ursula Renz, 1–36. Berlin, Boston: de Gruyter.
- Winko, Simone. 2003. *Kodierte Gefühle. Zu einer Poetik der Emotionen in lyrischen und poetologischen Texten um 1900*. Berlin: Erich Schmidt Verlag.
- Zehe, Albin, Martin Becker, Lena Hettinger, Andreas Hotho, Isabella Reger und Fotis Jannidis. 2016. Prediction of Happy Endings in German Novels based on Sentiment Information. In: *Proceedings of DMNLP*, hg. von Peggy Cellier, Thierry Charnois, Andreas Hotho, Stan Matwin, Marie-Francine Moens, und Yannick Toussaint, 9–17. Riva del Garda, Italy. <http://ceur-ws.org/Vol-1646/paper2.pdf>.
- Zhang, Lei und Bing Liu. 2014. Aspect and Entity Extraction for Opinion Mining. In: *Data Mining and Knowledge Discovery for Big Data*, hg. von Wesley W. Chu, 1–40. Heidelberg: Springer.
- Zhao, Yanchang, Huaifeng Zhang, Longbing Cao, Hans Bohlscheid, Yuming Ou und Chengqi Zhang. 2014. Data Mining Applications in Social Security. In: *Data Mining and Knowledge Discovery for Big Data*, hg. von Wesley W. Chu, 83–96. Heidelberg: Springer.

Glossar

- Annotation** Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch **Machine-Learning-Verfahren** durchgeführt wird. Ein klassisches Beispiel ist das automatisierte **PoS-Tagging** (Part-of-Speech-Tagging), welches oftmals als Grundlage (**Preprocessing**) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.
- Browser** Mit Browser ist in der Regel ein Webbrowser gemeint, also ein Computerprogramm, mit dem das Anschauen, Navigieren auf, und Interagieren mit Webseiten möglich wird. Am häufigsten genutzt werden dafür Chrome, Firefox, Safari oder der Internet Explorer.
- Close Reading** Close Reading bezeichnet die sorgfältige Lektüre und Interpretation eines einzelnen oder weniger Texte. Close Reading ist in der digitalen Literaturwissenschaft außerdem mit der manuellen **Annotation** textueller Phänomene verbunden (vgl. auch **Distant Reading** als Gegenbegriff).
- Commandline** Die Commandline (engl. *command line interface* (CLI)), auch Kommandozeile, Konsole, Terminal oder Eingabeaufforderung genannt, ist die direkteste Methode zur Interaktion eines Menschen mit einem Computer. Programme ohne eine grafische Benutzeroberfläche (**GUI**) werden i. d. R. durch Texteingabe in die Commandline gesteuert. Um die Commandline zu öffnen, klicken Sie auf Ihrem Mac „cmd“ + „space“, geben „Terminal“ ein und doppelklicken auf das Suchergebnis. Bei Windows klicken Sie die Windowstaste + „R“, geben „cmd.exe“ ein und klicken Enter.
- CSV** CSV ist die englische Abkürzung für *Comma Separated Values*. Es handelt sich um ein Dateiformat zur einheitlichen Darstellung und Speicherung von einfach strukturierten Daten mit dem Kürzel `.csv`, sodass diese problemlos zwischen IT-Systemen ausgetauscht werden können. Dabei sind alle Daten zeilenweise angeordnet. Alle Zeilen wiederum sind in einzelne Datenfelder aufgeteilt, welche durch Trennzeichen wie Semikola oder Kommata getrennt werden können. In Programmen wie Excel können solche Textdateien als Tabelle angezeigt werden.
- Data Mining** Data Mining gehört zum Fachbereich **Information Retrieval** und bezieht sich auf die systematische Anwendung computergestützter Methoden, die darauf abzielt, in vorhandenen Datenbeständen Muster, Trends oder Zusammenhänge zu erkennen. Textbasierte Formen des Data Minings sind u. a. **Text Mining**, **Web Mining** und **Opinion Mining**.
- Distant Reading** Distant Reading ist ein Ansatz aus den digitalen Literaturwissenschaften, bei dem computergestützte Verfahren auf häufig große Mengen an Textdaten angewandt werden, ohne dass die Texte selber gelesen werden. Meist stehen hier quantitative Analysen im Vordergrund, es lassen sich jedoch auch qualitative **Metadaten** quantitativ vergleichen. Als Gegenbegriff zu **Close Reading** wurde der Begriff insbesondere von Franco Moretti (2000) geprägt.
- Domäneadaption** Domäneadaption beschreibt die Anpassung einer in einem Fachgebiet entwickelten digitalen Methode an ein anderes Fachgebiet.
- GUI** GUI steht für *Graphical User Interface* und bezeichnet eine grafische Benutzeroberfläche. Ein GUI ermöglicht es, Tools mithilfe von grafischen Schaltflächen zu bedienen, um somit beispielsweise den Umgang mit der **Commandline** zu umgehen.
- HTML** HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von **Webbrowsern** dargestellt und geben

die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche **Metainformationen** enthalten, die auf einer Webseite selbst nicht ersichtlich sind.

- Information Retrieval** Die Teildisziplin der Informatik, das Information Retrieval, beschäftigt sich mit der computergestützten Suche und Erschließung komplexer Informationen in meist unstrukturierten Datensammlungen.
- Korpus** Ein Textkorpus ist eine Sammlung von Texten. Korpora (Plural für „das Korpus“) sind typischerweise nach Textsorte, Epoche, Sprache oder Autor*in zusammengestellt.
- Lemmatisieren** Die Lemmatisierung von Textdaten gehört zu den wichtigen **Preprocessing**-Schritten in der Textverarbeitung. Dabei werden alle Wörter (**Token**) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.
- Machine Learning** Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekanntem Daten verwendet werden.
- Markup Language** Markup Language bezeichnet eine maschinenlesbare Auszeichnungssprache, wie z. B. **HTML**, zur Formatierung und Gliederung von Texten und anderen Daten. So werden beispielsweise auch **Annotationen** durch ihre Digitalisierung oder ihre digitale Erstellung zu Markup, indem sie den Inhalt eines Dokumentes strukturieren.
- Metadaten** Metadaten oder Metainformationen sind strukturierte Daten, die andere Daten beschreiben. Dabei kann zwischen administrativen (z. B. Zugriffsrechte, Lizenzierung), deskriptiven (z. B. Textsorte), strukturellen (z. B. Absätze oder Kapitel eines Textes) und technischen (z. B. digitale Auflösung, Material) Metadaten unterschieden werden. Auch **Annotationen** bzw. **Markup** sind Metadaten, da sie Daten/Informationen sind, die den eigentlichen Textdaten hinzugefügt werden und Informationen über die Merkmale der beschriebenen Daten liefern.
- Named Entities** Eine Named Entity (NE) ist eine Entität, oft ein Eigenname, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.
- NLP** *Natural Language Processing* (NLP), maschinelle Sprachverarbeitung zu Deutsch, ist ein Teilgebiet der Linguistik, der Informatik und der künstlichen Intelligenz, welches sich damit beschäftigt, wie Computer so programmiert werden, dass sie große Mengen an natürlichsprachlichen Daten verarbeiten und analysieren können.
- Opinion Mining** Unter Opinion Mining, oder Sentiment Analysis, versteht man die Analyse von Stimmungen oder Haltungen gegenüber einem Thema, durch die Analyse natürlicher Sprache. Das Opinion Mining gehört zu den Verfahren des **Text Minings**.
- POS** PoS steht für *Part of Speech*, oder „Wortart“ auf Deutsch. Das PoS- **Tagging** beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist ein wichtiger **Preprocessing**-Schritt, beispielsweise für die Analyse von **Named Entities**.
- Preprocessing** Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden **lemmatisiert**.
- Reintext-Version** Die Reintext-Version ist die Version eines digitalen Textes oder einer Tabelle, in der keinerlei Formatierungen (Kursivierung, Metadatenauszeichnung etc.) enthalten sind. Reintext-Formate sind beispielsweise TXT, RTF und **CSV**.
- SBW** SBW steht für *Sentiment Bearing Word* (Sentimentwort) und bezeichnet ein Lexem, das eine Stimmung, eine Bewertung, ein Gefühl, eine Einstellung oder auch eine Emotion zum Ausdruck bringt. Für SBWs werden „semantische Orientierung“ und „Stärke“ in Form des **Sentimentwerts** festgelegt. SBWs werden in **Sentimentlexika** gesammelt und als Grundlage für lexikonbasierte Sentimentanalysen verwendet.
- Sentimentwert** Der Sentimentwert oder Sentimentgehalt eines Wortes beschreibt, meistens auf einer Skala von -1 (maximal negativ; bspw. „schädlich“: -0,9269) bis +1 (maximal positiv, bspw. „Freude“: 0,6502) die Polarität von Sentimentwörtern (siehe auch **SBWs**). Der Wert wird bei der Generierung von **Sentimentlexika** für jedes enthaltene Wort errechnet.
- Sentimentwörterbuch** Ein Wörterbuch, das ausschließlich Lexeme enthält, die **Träger von Sentiments** sind, wird als Sentimentwörterbuch definiert. Sentimentlexika fungieren als wichtige Ressource für lexikonbasierte Sentimentanalysen, bei denen die Wörter des Wörterbuchs und die Wörter eines Primärtextes miteinander abgeglichen werden.

Text Mining Das Text Mining ist eine textbasierte Form des **Data Minings**. Prozesse & Methoden, computer-gestützt und automatisch Informationen bzw. Wissen aus unstrukturierten Textdaten zu extrahieren, werden als Text Mining zusammengefasst.

Type/Token Das Begriffspaar „Type/Token“ wird grundsätzlich zur Unterscheidung von einzelnen Vorkommnissen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von Exemplaren (Token) umfasst.

Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär.“ beinhaltet beispielsweise fünf Worttoken („Ein“, „Bär“, „ist“, „ein“, „Bär“) und drei Types, nämlich: „ein“, „Bär“, „ist“. Allerdings könnten auch vier Types, „Ein“, „ein“, „Bär“ und „ist“, als solche identifiziert werden, wenn Großbuchstaben beachtet werden.

Unicode/UTF-8 Unicode ist ein internationaler Standard, der für jedes Schriftzeichen oder Textelement einen digitalen Code festlegt. Dabei ist UTF-8 die am weitesten verbreitete Kodierung für Unicode-Zeichen. UTF-8 ist die international standardisierte Kodierungsform elektronischer Zeichen und kann von den meisten Digital-Humanities-Tools verarbeitet werden.

Web Mining Unter Web Mining versteht man die Anwendung von Techniken des **Data Mining** zur Extraktion von Informationen aus dem World Wide Web. Das Web Mining ist ein Teilbereich des Data Minings und zählt zu einem der wichtigsten Anwendungsgebiete für das **Text Mining**.