Methodenbeitrag: Topic Modeling

Jan Horstmann [1]

1. Universität Münster

forTEXT

Thema:	Topic Modeling	DOI:	10.48694/fortext.3717
Jahrgang:	1	Ausgabe:	8
Erscheinungsdatum:	2024-07-10	Erstveröffentlichung:	2018-01-15 auf fortext.net
Lizenz:	CC BY SA		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

Topic Modeling ist ein auf Wahrscheinlichkeitsrechnung basierendes Verfahren zur Exploration (vgl. Text Mining) größerer Textsammlungen (vgl. Korpus). Das Verfahren erzeugt statistische Modelle (Topics) zur Abbildung häufiger gemeinsamer Vorkommnisse von Wörtern.

Abb. 1: Beispiele für Wörter des Topics „Theater“

1. Definition

Die Methode des Topic Modeling bietet die Möglichkeit, Textsammlungen thematisch zu explorieren. Dabei geht man davon aus, dass eine Textsammlung aus unterschiedlichen ‚Themen' bzw. besser: ‚Topics' besteht, die in den einzelnen Dokumenten der Sammlung in unterschiedlicher Ausprägung vertreten sind. Unter einem

‚Topic' versteht man dabei eine Gruppe von Wörtern (wie zum Beispiel die Wörter „Theater", „Schauspieler" und „Stück"), die in einem Text ungewöhnlich – d. h. statistisch auffällig – oft gemeinsam vorkommen. Ein ‚Topic' ist also ein statistisches Phänomen und damit zwar eine Entsprechung, aber nicht exakt das Gleiche wie ein (inhaltlich definiertes) Thema.

2. Anwendungsbeispiel

Angenommen, Sie möchten den Inhalt einer größeren Textsammlung – wie beispielsweise das Œuvre Therese Hubers oder auch die gesamte Prosaliteratur des 18. und 19. Jahrhunderts – erforschen. Digitale Verfahren können Sie dabei mit Methoden unterstützen, die dem Distant Reading zugeordnet werden. Ohne jeden Text der Sammlung individuell zu lesen, wird es dadurch möglich, die Texte untereinander zu vergleichen. Abhängig von der Größe Ihrer Textsammlung können Sie selbst entscheiden, wie viele ‚Topics' erstellt werden und wie groß diese Topics sein sollen. Als Nutzer*in der Methode modellieren Sie die Topics so lange, bis sie Ihnen aussagekräftig genug erscheinen, um anschließend zu untersuchen, welche Topics in welchen Texten besonders oft vertreten sind – oder auch umgekehrt, welche Texte ein gegebenes Topic besonders stark ‚thematisieren'.

3. Literaturwissenschaftliche Tradition

In seinem Beitrag zur Inhaltsanalyse im Handbuch Literaturwissenschaft konstatiert Anz (2007, 57): „Eine Beschäftigung vor allem mit Textinhalten setzt sich in der Literaturwissenschaft dem topischen Vorwurf oder Verdacht aus, die Formen von Texten zu vernachlässigen". Die Betrachtung der Interdependenzen von Inhalt und Form bildet daher nicht selten den Kern von Analysen beispielsweise der Literarizität von Texten.
Bei der Erschließung eines zu erforschenden Textes ist jedoch häufig der erste Schritt, sich einen Überblick über die im Text behandelten Themen zu verschaffen, d. h. nach einem „Leitgedanken [zu suchen], nach dem sich [sein] Inhalt zusammenfassen läßt" (Schulz 2003, 634), oder auch nach der „abstrakte[n] Grundkonstellation, die in Darstellung und Geschehen konkret ausgestaltet wird" (ebd.). Durch den Vergleich von in der Literatur wiederkehrenden Themen wird es möglich, „sowohl Rückschlüsse auf den Vorgang der menschlichen Orientierung im Dasein als auch auf die in ihm zum Ausdruck kommenden geistesgeschichtlichen Umschichtungen einer Zeit" (Daemmrich und Daemmrich 1995) zu ziehen.
Die den Topics etymologisch näheren literarischen Topoi bezeichnen seit Curtius (1948) solche literarischen Gemeinplätze, die aufgrund ihres hohen Alters „zwischen Altehrwürdigkeit und Abgegriffenheit" (Müller 2004, 279) schwanken. Oft ist es auch die thematische Schwerpunktsetzung, die einen Vergleich unterschiedlicher Texte des gleichen Autors oder verschiedener Autoren initiiert. Die „Stoff- und Motivgeschichte" bzw. die „Thematologie" wird daher auch als „Teildisziplin der Komparatistik" (Lubkoll 2004, 255) bezeichnet. Nicht zuletzt liegt jeder Literaturgeschichtsschreibung (a) eine große Menge an Texten und (b) die nachgewiesene Kenntnis der Inhalte dieser Texte zugrunde (Anz 2007, 55). Auch eher kulturwissenschaftlich ausgerichteten literaturwissenschaftlichen Arbeiten geht häufig eine Orientierung auf motiv- oder themengeschichtliche Zusammenhänge von literarischen und nicht-literarischen Texten voraus.
Als Literaturwissenschaftler*innen dient uns bei der Auswahl thematisch relevanter Texte für eine Fragestellung bislang häufig die eigene Forschungshistorie oder das angeeignete Fachwissen über mehr oder weniger kanonisierte Texte. Die Methode des Topic Modeling eignet sich zunächst gut, um große Textsammlungen zu explorieren, gleichzeitig bilden jedoch auch literaturwissenschaftliche Kenntnisse über zumindest eine Teilmenge der analysierten Texte bzw. die Art und Weise der Behandlung bestimmter stofflicher Phänomene durch eine Autorin wichtige Grundbedingungen, um die entstehenden Topics interpretieren zu können.

4. Diskussion

Gerade bei größeren Textsammlungen wie der Prosaliteratur des 19. Jahrhunderts oder auch umfangreichen Texten wie z. B. Prousts *Recherche* werden Sie als Literaturwissenschaftler*in häufig nicht die Kapazitäten haben, sämtliche Texte detailliert zu lesen bzw. zu analysieren (vgl. Close Reading). Zusätzlich ist es dem menschlichen Gehirn nicht möglich, Textmengen ab einer bestimmten Größe gleichzeitig zu überschauen und insgesamt miteinander zu vergleichen. Die Methode verspricht durch die Fokussierung auf die Thematik, sich den semantischen Strukturen der analysierten Texte zu nähern – wodurch sie sich von rein quantitativen DH-Methoden unterscheidet. Betont werden sollte dabei, dass die resultierenden Topics nicht selbst die Semantik der Texte abbilden, sondern dass textimmanente Bedeutungsstrukturen in ihnen abgelesen werden können. Literaturwissenschaftliches Fachwissen ist bei der Auswertung daher unumgänglich, weshalb auch die Topic-Modeling-Exploration eines mittelgroßen Korpus viele Vorteile bietet (Weitin und Herget 2016, 3f.).
Jannidis (2016, 27) beobachtet: „Schon früh ist den Fachwissenschaftlern, die mit Computerlinguisten und Informatikern an Topic-Modeling-Projekten arbeiten, aufgefallen, dass auch Worte, die aufgrund von bestimmten rhetorischen Strukturen auftauchen, als ‚Thema' zusammengefasst wurden". Diese rhetorischen Strukturen gehen jedoch schnell verloren, wenn man beispielsweise ausschließlich Topics aus Substantiven bildet, wie

Jockers (2013) es durchführt.
„Topics“ sollten zudem nicht mit literarischen „Themen“ gleichgesetzt werden. Während Topics Häufigkeiten und Verteilungen ausschließlich auf der Wortoberfläche abbilden, können Themen auch implizit sein: Das virulente Thema der Homosexualität in Prousts *Recherche* wird als solches beispielsweise nie direkt adressiert, geschweige denn benannt. Topics sind daher für sich keine Themen, können jedoch als solche interpretiert werden, wodurch der Methode der Charakter einer textanalytischen Heuristik zugesprochen werden kann. Topics sind daher weniger ‚Themen' als vielmehr ein Indikator für den jeweils verarbeiteten literarischen Stoff: „Anders als Stoff bezeichnet Thema nicht das konkrete, an Figurenkonstellationen und Handlungszüge gebundene Material, das in einem Text verarbeitet wird, sondern die darin enthaltene Problemkonstellation: ‚Romeo und Julia' (Stoff) vs. ‚illegitime Liebesbeziehung' (Thema, aber auch Motiv)“ (Schulz 2003, 634). Diese Abgrenzung ziehen wir jedoch auch in der Literaturwissenschaft selbst nicht immer strikt: Daemmrich und Daemmrich (Daemmrich und Daemmrich 1995, XIII) sprechen von der „Tendenz, die Kategorie [Stoff] zu erweitern und sie anderen Begriffen wie Sujet, Topos, Motiv, Mythos und Thema anzugleichen“. Die dem Stoff implizit eingeschriebenen Themen müssen Sie im Zuge der literaturwissenschaftlichen Auslegung der erhaltenen Topics feststellen.

5. Technische Grundlagen

Der im Topic Modeling am häufigsten genutzte Algorithmus wurde von Blei, Ng und Jordan (2003) unter dem Namen Latent Dirichlet Allocation (LDA) entwickelt (Blei 2012) und liegt auch dem Tool Mallet zugrunde. Er basiert auf einer wiederholt zufälligen Auswahl an Textsegmenten, wobei innerhalb dieser Segmente jeweils die statistische Häufung von Wortgruppen erfasst wird. Der Algorithmus berechnet somit die Topics der Textsammlung, die Topic-Anteile in den Einzeltexten und welche Wörter zu den jeweiligen Topics gehören.
Als Nutzer*in können Sie die Menge und Größe der zu erstellenden Topics sowie die Anzahl der Iterationen bestimmen. Mallet können Sie beispielsweise in der Software R nutzen; Ihnen wird hier jedoch keine grafische Nutzeroberfläche (vgl. GUI) geboten, sodass grundlegende Kenntnisse im Coding (vgl. CODE) vonnöten sind, um die Texte vorzubereiten, dann das Topic Modeling selbst durchzuführen und schließlich die Ergebnisse auszuwerten und zu visualisieren. Besonders hilfreich ist hier die für die Bedarfe und Horizonte von Geisteswissenschaftler*innen zugeschnittene Einführung von Jockers (2014), die auch ein Kapitel zum Topic Modeling enthält. Für den Einstieg bietet sich die Arbeit mit dem DARIAH Topics Explorer (Schumacher 2024a) an, in dem Sie Topics mithilfe einer grafischen Nutzeroberfläche modellieren können.
Topic Modeling ist ein probabilistisches, unüberwachtes Verfahren (vgl. Machine Learning), d. h. Sie können zwar die genannten Parameter bestimmen und die Ergebnisse analysieren, in den automatischen Prozess der Modellierung selbst haben Sie jedoch keinen direkten Einblick und die Textsegmentauswahl erfolgt zufällig. Da die Ergebnisse des Topic Modelings auf komplexen Wahrscheinlichkeitsberechnungen basieren, ist ein Topic Modeling – auch wenn Sie die wählbaren Parameter (vgl. Hyperparameter) gleich einstellen – nicht eins zu eins reproduzierbar, wenn auch eine große Ähnlichkeit unter den entstehenden Topics zu erkennen ist. Außerdem macht das Verfahren Gebrauch von einer Stoppwortliste, die für gewöhnlich die in Texten am häufigsten verwendeten, für sich genommen jedoch selten einen eigenen semantischen Wert aufweisenden Wörter (MFW = *most frequent words*) enthält. Die Stoppwortliste erweitern Sie nach jedem vollständigen Durchlauf um diejenigen Wörter, die in den resultierenden Topics auftauchen, jedoch keinen Erkenntnismehrwert bringen.

Um die Ergebnisse zu verfeinern, können Sie im Zuge des Preprocessing der Texte außerdem mehrere Aktionen durchführen:

1. Um eine getrennte Behandlung von (am Satzanfang) groß und (innerhalb des Satzes) klein geschriebenen Varianten desselben Wortes zu vermeiden, wandelt man in der Regel sämtliche Buchstaben in Kleinbuchstaben um.
2. Eine Lemmatisierung (vgl. Lemmatisieren) bewirkt, dass Varianten eines Wortes auf ihre Grundform (Lemma) reduziert und folglich als gleiches Wort behandelt werden können.
3. Ein *part of speech*-Tagging (POS-Tagging) ermöglicht Ihnen die getrennte Untersuchung von ausgewählten Wortgruppen. Einige Forscher betreiben Topic Modeling beispielsweise ausschließlich mit Substantiven (Jockers 2013).
4. Eine Annotation der Named Entities ermöglicht es Ihnen, alle Eigennamen gebündelt aus dem Topic Modeling auszuschließen. Alternativ müssen Sie die in den Topics auftauchenden Eigennamen nach jedem Durchgang auf die Stoppwortliste setzen, wenn sie nicht in den Ergebnissen vertreten sein sollen – zur Problematik von Eigennamen im Topic Modeling (Jockers 2013). (Mehr zur Named Entity Recognition bei Schumacher (2024b))

Sind Sie mit dem Ergebnis der entstehenden Topics in Umfang und Genauigkeit zufrieden, haben Sie unterschiedliche Möglichkeiten der Visualisierung (vgl. Textvisualisierung (Horstmann und Stange 2024)): Topics werden zunächst als Wortliste herausgegeben, die sich in R aber beispielsweise auch als Wordclouds darstellen lassen. Um die eigene Textsammlung zu explorieren, bietet es sich an, für alle oder ausgewählte Topics Balkendiagramme erstellen zu lassen, die anzeigen, wie häufig das jeweilige Topic in den einzelnen Dokumenten der

Textsammlung vorkommt (‚documents per topic'). Interessieren Sie sich für bestimmte Texte der Sammlung, lässt sich ebenso anzeigen, wie häufig die einzelnen Topics in den jeweiligen Texten vorkommen (‚topics per document').

Externe und weiterführende Links

- Mallet: https://web.archive.org/save/http://mallet.cs.umass.edu/topics.php (Letzter Zugriff: 22.08.2024)
- Software R: https://web.archive.org/save/https://www.r-project.org (Letzter Zugriff: 28.07.2024)

Bibliographie

Anz, Thomas, Hrsg. 2007. Inhaltsanalyse. In: *Handbuch Literaturwissenschaft*, 2: Methoden und Theorien:55–69. Stuttgart, Weimar: Metzler.

Blei, David M. 2012. Probabilistic topic models. *Communications of the ACM* 55, Nr. 4 (April): 77–84. doi: 10.1145/2133806.2133826, https://dl.acm.org/doi/10.1145/2133806.2133826 (zugegriffen: 14. Juli 2020).

Blei, David M, Andrew Y Ng und Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, Nr. Jan: 993–1022.

Curtius, Ernst Robert. 1948. *Europäische Literatur und lateinisches Mittelalter*. Bern: Francke.

Daemmrich, Horst S. und Ingrid G. Daemmrich. 1995. *Themen und Motive in der Literatur. Ein Handbuch*. Tübingen, Basel: Francke.

Horstmann, Jan und Jan-Erik Stange. 2024. Methodenbeitrag: Textvisualisierung. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 5. Textvisualisierung (7. August). doi: 10.48694/fortext.3772, https://fortext.net/routinen/methoden/textvisualisierung.

Jannidis, Fotis. 2016. Quantitative Analyse literarischer Texte am Beispiel des Topic Modeling. *Der Deutschunterricht* 68, Nr. 5: 24–35.

Jockers, Matthew. 2013. „Secret" Recipe for Topic Modeling Themes. *Matthew L. Jockers*. http://www.matthewjockers.net/2013/04/12/secret-recipe-for-topic-modeling-themes/ (zugegriffen: 24. November 2017).

———. 2014. *Text Analysis With R for Students of Literature*. Cham (u.a.): Springer.

Lubkoll, Christine. 2004. Stoff- und Motivgeschichte/Thematologie. In: *Grundbegriffe der Literaturtheorie*, hg. von Ansgar Nünning, 255–259. Stuttgart, Weimar: Metzler.

Müller, Wolfgang G. 2004. Topik/Toposforschung. In: *Grundbegriffe der Literaturtheorie*, hg. von Ansgar Nünning, 278–280. Stuttgart, Weimar: Metzler.

Schulz, Armin. 2003. Thema. In: *Reallexikon der deutschen Literaturwissenschaft. Neubearbeitung des Reallexikons der deutschen Literaturgeschichte*, 3: P-Z:634–635. Berlin, New York: de Gruyter.

Schumacher, Mareike. 2024a. Toolbeitrag: DARIAH Topics Explorer. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 8. Topic Modeling (7. Oktober). doi: 10.48694/fortext.3728, https://fortext.net/tools/tools/dariah-topics-explorer.

———. 2024b. Methodenbeitrag: Named Entity Recognition (NER). Hg. von Evelyn Gius. *forTEXT* 1, Nr. 9. Named Entity Recognition (30. Oktober). doi: 10.48694/fortext.3765, https://fortext.net/routinen/methoden/named-entity-recognition-ner.

Weitin, Thomas und Katharina Herget. 2016. Falkentopics. *LitLab Pamphlet #4*. http://www.digitalhumanitiescooperation.de/wp-content/uploads/2017/06/p04_weitin_herget_de.pdf (zugegriffen: 24. November 2017).

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).

CODE Der Code, oder auch Programmcode/ Maschinencode, bezieht sich auf eine Sammlung von Anweisungen, die durch verschiedene Programmiersprachen wie Java, Python oder C realisiert werden können. Für die Ausführung der Anweisungen wird der Code durch einen Compiler oder einen Interpreter in die Maschinensprache, einen Binärcode, des Computers übersetzt.

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.

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 computationelle 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.

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.

Hyperparameter Hyperparameter beziehen sich auf externe, anpassbare Einstellungen, die genutzt werden um den Lernprozess zu kontrollieren und zu beeinflussen (zu modellinternen Parametern siehe Parameter). Sie sind unabhängig vom Datensatz und beziehen sich beispielsweise auf Einstellungen wie Anzahl der Iterationen, Größe der Datensätze oder Kontextfenster.

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.

LDA LDA steht für *Latent Dirichlet Allocation* und ist ein generatives, statistisches Wahrscheinlichkeitsmodell, welches zum Topic Modeling angewendet werden kann. Bei der LDA werden auf Grundlage eines Wahrscheinlichkeitsmodells Wortgruppen aus Textdokumenten erstellt. Dabei wird jedes Dokument als eine Mischung von verborgenen Themen betrachtet und jedes Wort einem Thema zugeordnet. Wortreihenfolgen und Satzzusammenhänge spielen dabei keine Rolle.

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 unbekannten 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.

Opinion Mininig 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.

Parameter Im Kontext von Machine-Learning-Modellen handelt es sich bei (Modell-)Parametern um modellinterne Konfigurationsvariablen, die anhand des Trainingssatzes bestimmt werden (zu modellexternen Parametern siehe Hyperparameter). Als Parameter werden einerseits Aspekte benannt, die den Lernprozess bestimmen und andererseits solche, die dabei erlernt werden. Die Werte der Parameter ergeben sich aus dem Datensatz selbst. Werte solcher Parameter können beispielsweise die Gewichtungen in neuronalen Netzwerken sein, also welche Aspekte im Trainingsprozess besonders einflussreich sind (z. B. können Wörter im direkten Umfeld eines Zielwortes als wichtiger bewertet werden also solche, die weit von diesem entfernt stehen) oder etwa wie die Gewichtung (also die Reihenfolge) der einzelnen Wörter innerhalb der Topics beim Topic Modeling.

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 of 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.

Stoppwortliste Stoppwörter sind hochfrequente Wörter, meist Funktionswörter, die, aufgrund ihrer grammatisch bedingten Häufigkeit, beispielsweise die Ergebnisse von inhaltlichen oder thematischen Analysen verzerren können. Deshalb werden diese Wörter, gesammelt in einer Stoppwortliste, bei digitalen Textanalysen meist nicht berücksichtigt.

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

Topic Modeling Das Topic Modeling ist ein statistisches, auf Wahrscheinlichkeitsrechnung basierendes, Verfahren zur thematischen Exploration größerer Textsammlungen. Das Verfahren erzeugt „Topics" zur Abbildung häufig gemeinsam vorkommender Wörter in einem Text. Für die Durchführung können verschiedene Algorithmen und Modelle wie das LDA verwendet werden.

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.

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.

Wordcloud Eine *Wordcloud*, oder auch Schlagwortwolke, ist eine Form der Informationsvisualisierung, beispielsweise von Worthäufigkeiten in einem Text oder einer Textsammlung. Dabei werden unterschiedlich gewichtete Wörter, wie die häufigsten Wörter, i.d.R. größer oder auf andere Weise hervorgehoben dargestellt. Die horizontale/vertikale Ausrichtung und die Farbe der dargestellten Wörter hat meistens allerdings keinen semantischen Mehrwert.