Toolbeitrag: LIWC

Marie Flüh [1]

1. Universität Hamburg

forTEXT

Thema:	Sentimentanalyse	DOI:	10.48694/fortext.3800
Jahrgang:	1	Ausgabe:	7
Erscheinungsdatum:	2024-07-10	Erstveröffentlichung:	2019-08-12 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.

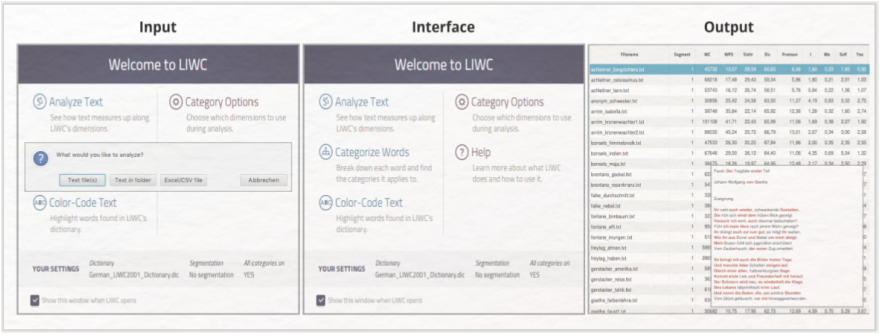

Abb. 1: Der Workflow von LIWC: Vorab: Installation des Tools, Download des deutschsprachigen Lexikons (optional); Input: Hochladen einzelner oder mehrerer Texte; Interface: Auswahl der gewünschten Analyseform; Output: Ansicht der Analyseergebnisse und Export

- **Systemanforderungen:** Die kostenfreie Demoversion ist für eine maximale Textgröße von 1000 Wörtern ausschließlich webbasiert (vgl. Browser) nutzbar, die kostenpflichtige LIWC-Version steht für Windows- und Mac-Betriebssysteme zur Verfügung (Ausnahme: nicht kompatibel mit Windows XP und Macintosh OSX10.7.x – Lion)
- **Stand der Entwicklung:** In den 1990er Jahren entwickelt, 2001 erstveröffentlicht, aktuelle Version: LIWC2015 v1.6
- **Herausgeber:** LIWC Inc. (James W. Pennebaker, Roger J. Booth, Martha E. Francis)
- **Lizenz:** Kommerzielle Version mit unterschiedlicher Laufzeit bzw. variierendem Funktionalitätsumfang (LIWCLITE 7, LIWC 2007, LIWC 2015), Lizenzmodelle mit Vergünstigungen für die Verwendung in Bildung und Forschung, Nutzung der LIWC-API (vgl. API) nach erfolgreicher Bewerbung und genauen Angaben zum Verwendungszweck
- **Weblink:** https://www.liwc.app
- **Im- und Export:** Import von Dateien in den Formaten DOCX, DOC, TXT, RTF (vgl. Reintext-Version), PDF, XLSX, XLS, CSV; Export der Ergebnisse in diversen und mit SPSS, R, SAS, SciPy oder Weka kompatiblen Formaten wie TXT, CSV oder als Excel-Datei
- **Sprachen:** Arabisch, Chinesisch, Niederländisch, Englisch, Deutsch, Französisch, Italienisch, Portugiesisch, Russisch, Serbisch, Spanisch und Türkisch

1. Für welche Fragestellungen kann LIWC eingesetzt werden?

Ursprünglich entwickelt, um Essays aus Experimenten zum expressiven bzw. therapeutischen Schreiben zu untersuchen (Wolf u. a. 2008), eignet sich LIWC („Linguistic Inquiry and Word Count“) für die Analyse diverser Textsorten wie persönlichen, subjektiven Texten, E-Mail-Korrespondenzen, Social-Media-Beiträgen wie Tweets oder Blogeinträgen, Werbetexten oder wissenschaftlichen Texten. Das Tool wurde in unterschiedlichen Studien zu persönlichkeits-, sozial- und klinisch-psychologischen Fragestellungen und für die Analyse von therapeutischen Essays, Alltagskommunikation, computervermittelter Kommunikation (Vergani und Bliuc 2015; Back, Küfner und Egloff 2011), (politischen) Reden (Abe 2011) sowie genderspezifischer Sprache (Newman u. a. 2008) eingesetzt und gilt als zuverlässiges Softwareprogramm zur quantitativen Textanalyse (vgl. Distant Reading) (Hai-Jew 2016; Wolf u. a. 2008; Proyer und Brauer 2018; Pennebaker und Chung 2008). Literaturwissenschaftlich relevante Fragestellungen, die Sie mit LIWC untersuchen können, betreffen die textbasierte Erforschung der emotionalen Dimension literarischer Texte wie beispielsweise: Welche emotionalen Affekte (wie Angst oder

Aggressivität) prägen Edgar Allan Poes Kurzgeschichte *The Tell-Tale Heart*? Überwiegen die positiven oder die negativen Emotionen in Lewis Carrolls *Alice in Wonderland* und wie lässt sich die emotionale Tonalität der Novelle beschreiben?

2. Welche Funktionalitäten bietet LIWC und wie zuverlässig ist das Tool?

Die Konzeption des Tools basiert auf der Grundannahme, dass sich die Persönlichkeit des Menschen in der Sprache widerspiegelt, die er verwendet. LIWC liegt die auf gesprächstherapeutischer Forschung basierende Annahme zugrunde, dass die Analye der verwendeten Funktionswörter, die in der Kommunikation zumeist unbewusst eingesetzt werden (wie z. B. Pronomen, Artikel und Konjunktionen), besonders aussagekräftig ist. Rückschlüsse auf sich im Inneren des Verfassers abspielende Prozesse lassen sich folglich u. a. durch die Analyse der „kleinen" Wörter ziehen. Die Verwendung von Inhaltswörtern (Substantive, Adjektive und Verben), die zwar die Bedeutung eines Satzes tragen, aber deutlich stärker von externen Faktoren (wie der Vorgabe eines bestimmten Themas) beeinflusst werden, spielt bei der Analyse eine sekundäre Rolle. Inhaltliche Zusammenhänge werden bei der Textanalyse mit LIWC gänzlich ausgeblendet.
LIWC führt eine automatisierte Ein-Wort-Analyse (vgl. Text Mining); (*word-by-word basis*) auf Basis eines v. a. von Psycholog*innen entworfenen Lexikons durch. Sobald Sie im Besitz einer kommerziellen LIWC-Version sind, können Sie die Wörterbücher einsehen, insofern Sie die Demoversion verwenden, stehen diese jedoch nicht zur freien Verfügung. Es gilt zu bedenken, dass sich alltägliche Sprache und innerliterarische, innerkünstlerische Sprache erheblich unterscheiden. Eine LIWC-basierte Analyse eines literarischen Textes – unter Rückgriff auf ein eher psychophysisch, alltagssprachlich ausgerichtetes Wörterbuch – ist fragwürdig: Emotivität ist in literarischen Texten auf andere Art und Weise kodiert als anderen Textgattungen. Darüber hinaus finden sich Emotionen als textuelle Phänomene nicht nur auf allen sprachlichen Ebenen (Morpheme, Wörter, Sätze). Die Informationsstruktur des gesamten Textes ist emotionskonstituierend und -ausdrückend, literarische Texte weisen unterschiedliche emotionale Dimensionen auf (Schwarz-Friesel 2017), die durch eine Ein-Wort-Analyse kaum erfasst werden können. Nach dem Erwerb einer lizenzierten Version können Sie allerdings nicht nur das deutschsprachige LIWC-Lexikon herunterladen, sondern auch eigens erstellte Lexika integrieren. Bei der Konzeption (in Word oder Excel) und Implementierung (als .txt-Datei, die allerdings auf .dic endend abgespeichert werden muss) müssen Sie die LIWC-Syntax beachten. Hierbei können Sie z. T. reguläre Ausdrücke (vgl. Reguläre Ausdrücke) verwenden, was die Erstellung eines umfassenden textsortenspezifischen Lexikons erleichtert (indem Sie z. B. sämtliche Flexionsformen eines Wortes durch ein * am Ende des Stammwortes abfragen (vgl. Query)).
In der Standardeinstellung greift LIWC auf das integrierte englischsprachige LIWC-Lexikon aus dem Jahr 2015 zurück. Bei Bedarf können Versionen aus den Jahren 2001 und 2007 aktiviert werden. Darüber hinaus wurde das LIWC-Lexikon nicht nur in die deutsche, sondern auch in diverse weitere Sprachen (italienisch, norwegisch, spanisch, brasilianisch, portugiesisch, französisch, niederländisch, russisch, traditionelles wie vereinfachtes Chinesisch) übertragen. Für die Mehrzahl der LIWC-Kategorien kann eine gute Äquivalenz der deutschen Version mit dem englischen Original bestätigt werden, für einige basislinguistischen Kategorien wurden allerdings Unterschiede festgestellt (Wolf u. a. 2008).

Funktionen:

Bei der Verwendung der ausschließlich mit englischsprachigen Texten funktionierenden **Demoversion** u. a.:

- Automatisierte Ein-Wort-Analyse (*word-by-word basis*, vgl. Type/Token): Im Kern nutzt dieses Analyseverfahren einen Wortzählalgorithmus, der die Wörter eines Textes auszählt und diese vorab definierten und in einem internen Wörterbuch organisierten Wortkategorien zuordnet (Wolf u. a. 2008). Die Analyse spielt sich folglich ausschließlich auf der lexikalischen Ebene ab und basiert auf dem Abgleich des hochgeladenen individuellen Textes mit dem implementierten LIWC-Lexikon.
- Auskunft über prozentualen Anteil der *I-Words* (I, me, my), der *Social Words*, der *Positive Emotions*, der *Negative Emotions* und der *Cognitive Processes*.
- Die Variable *Analytical Thinking* erfasst den Grad, in dem die schreibende Person Wörter verwendet, die auf formale, logische und hierarchische Denkstrukturen verweisen.
- *Clout* beschreibt den sozialen Status bzw. das Selbstbewusstsein und Führungsverhalten, das die schreibende Person zum Ausdruck bringt.
- *Authenticity* erfasst, ob die schreibende Person authentisch und ehrlich kommuniziert.
- *Emotional Tone* erfasst, ob dem untersuchten Dokument ein überwiegend positiver oder negativer Ton zugrunde liegt.
- Darüber hinaus werden vergleichende Daten zur Verfügung gestellt, die zeigen, wie Texte derselben Kategorie durchschnittlich zusammengesetzt sind.

Die **kommerzielle Variante** des Tools beinhaltet drei Darstellungsweisen der Ein-Wort-Analyse:

- *Analyze Text* ist eine tabellarische Übersicht der Analyseergebnisse des gesamten Dokuments (Dokumentebene).
- *Categorize Text* bietet eine Liste sämtlicher Wörter mit Angabe der jeweiligen Kategorie (Wortebene).
- *Color-Code Text* ist eine Ansicht des gesamten Textes bei farblicher Hervorhebung derjenigen Wörter, die einer Kategorie zugeordnet wurden (Satzebene).

Zuverlässigkeit: LIWC funktioniert zuverlässig. Sofern Sie keine einzelnen Texte, sondern ein größeres Textkorpus (vgl. Korpus) untersuchen möchten, kann der Analyseprozess jedoch einige Zeit dauern.

3. Ist LIWC für DH-Einsteiger*innen geeignet?

Checkliste	√ / teilweise / –
Methodische Nähe zur traditionellen Literaturwissenschaft	teilweise
Grafische Benutzeroberfläche	√
Intuitive Bedienbarkeit	√
Leichter Einstieg	√
Handbuch vorhanden	√
Handbuch aktuell	√
Tutorials vorhanden	√
Erklärung von Fachbegriffen	√
Gibt es eine gute Nutzerbetreuung?	√

Detaillierte Erklärungen der Funktionalitäten finden Sie auf der LIWC-Homepage. Darüber hinaus helfen Ihnen Tutorials dabei, LIWC schrittweise kennenzulernen und unterschiedliche Funktionen – wie z. B. die Konzeption eines individuellen Lexikons – auszuführen. Relevante Funktionen lassen sich dank einer intuitiv bedienbaren Benutzeroberfläche (vgl. GUI) aber auch ohne die Konsultation eines Handbuchs und technisches Vorwissen ausführen. Nutzeranfragen per E-Mail werden zuverlässig und in kurzer Zeit beantwortet.

4. Wie etabliert ist LIWC in den (Literatur-)Wissenschaften?

In seinem ursprünglichen Forschungsbereich – der Psychologie – ist das Tool etabliert, auch wenn es aufgrund des Außerachtlassens komplexerer Bedeutungsstrukturen durchaus kontrovers diskutiert wird. In der Literaturwissenschaft wurde das Tool bisher kaum verwendet, obwohl es sich durch die Möglichkeit, ein individuelles deutschsprachiges Lexikon zu integrieren, durchaus für die Analyse deutschsprachiger Texte eignet und Untersuchungen zufolge bspw. für die Analyse lyrischer Texte oder Erzählungen in Frage käme (Wolf u. a. 2008). Eine literaturwissenschaftliche Adaption (vgl. Domäneadaption) wäre außerdem möglich, da sich die intuitiv bedienbare GUI von LIWC mit der Verwendung von existierenden Sentimentwörterbüchern wie SentiWS oder – besser noch – domänenspezifischen, eigens entworfenen Sentimentwörterbüchern (Sentimentanalyse (Flüh 2024)), die z. B. historische und orthographische Besonderheiten einbeziehen, kombinieren ließe. Es gilt festzuhalten, dass für die Analyse deutsprachiger literarischer Texte ein Lexikon benötigt wird, welches durch spezifische Analysekategorien der Beschaffenheit literarischer Texte gerecht wird.

5. Unterstützt LIWC kollaboratives Arbeiten?

Nein, LIWC ist für die Einzelarbeit konzipiert.

6. Sind meine Daten bei LIWC sicher?

Ja, sobald Sie eine LIWC-Version erworben haben, wird das Tool desktopbasiert ausgeführt. Hier sind Ihre Textdaten sicher. Zur Zahlungsabwicklung müssen Sie personenbezogene Daten angeben, was sich bei der Verwendung der Demoversion erübrigt.

Externe und weiterführende Links

- LIWC: https://web.archive.org/save/https://www.liwc.app (Letzter Zugriff: 17.09.2024)
- Konzeption eines individuellen Lexikons: https://web.archive.org/save/https://www.youtube.com/watch?v=CXPfrkfs7eo (Letzter Zugriff: 17.09.2024)

- SentiWS: https://web.archive.org/save/http://wortschatz.uni-leipzig.de/de/download (Letzter Zugriff: 28.07.2024)

Bibliographie

Abe, Jo Ann A. 2011. Changes in Alan Greenspan's Language Use Across the Economic Cycle: A Text Analysis of His Testimonies and Speeches. *Journal of Language and Social Psychology* 30, Nr. 2: 212–223. doi: 10.1177/0261927X10397152, (zugegriffen: 1. Juli 2019).

Back, Mitja D., Albrecht C. P. Küfner und Boris Egloff. 2011. Automatic or the People? Anger on September 11, 2001, and Lessons Learned for the Analysis of Large Digital Data Sets. *Psychological Science* 22, Nr. 6: 837–838. doi: 10.1177/0956797611409592, (zugegriffen: 30. Juni 2019).

Flüh, Marie. 2024. Methodenbeitrag: Sentimentanalyse. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 7. Sentimentanalyse (7. Oktober). doi: 10.48694/fortext.3797, https://fortext.net/routinen/methoden/sentimentanalyse.

Hai-Jew, Shalin. 2016. Extracting Linguistic Patterns from Texts with LIWC („luke") for Analysis. *C2C Digital Magazine (Fall 2016 / Winter 2017)*. https://scalar.usc.edu/works/c2c-digital-magazine-fall-2016--winter-2017/extracting-linguistic-patterns-from-texts-liwc-analysis (zugegriffen: 1. Juli 2019).

Newman, Matthew L., Carla J. Groom, Lori D. Handelman und James W. Pennebaker. 2008. Gender Differences in Language Use: An Analysis of 14,000 Text Samples. *Discourse Processes* 45, Nr. 3: 211–236. doi: 10.1080/01638530802073712, (zugegriffen: 1. Juli 2019).

Pennebaker, James W. und Cindy K. Chung. 2008. Computerized Text Analysis of Al-Qaeda Transcripts. In: *The content analysis reader*, hg. von Klaus Krippendorf und Mary Angela Bock, 453–467. Los Angeles (u.a.): SAGE Publications.

Proyer, René T. und Kay Brauer. 2018. Exploring adult playfulness: Examining the accuracy of personality judgments at zero-acquaintance and an LIWC analysis of textual information. *Journal of Research in Personality* 73: 12–20. doi: 10.1016/j.jrp.2017.10.002, (zugegriffen: 1. Juli 2019).

Schwarz-Friesel, Monika. 2017. Das Emotionspotenzial literarischer Texte. In: *Handbuch Sprache in der Literatur*, hg. von Anne Betten, Ulla Fix, und Berbelin Wanning, 17:351–370. Berlin, Boston: de Gruyter.

Vergani, Matteo und Ana-Maria Bliuc. 2015. The evolution of the ISIS' language: A quantitative analysis of the language of the first year of Dabiq magazine. *Sicurezza, terrorismo e società* 2: 7–20.

Wolf, Markus, Andrea Mehl, Matthias Severin, Haug Severin, James W. Pennebaker und Hans Kordy. 2008. Computergestützte quantitative Textanalyse: Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count. *Diagnostica* 54, Nr. 2: 85–98. doi: 10.1026/0012-1924.54.2.85,.

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.

API API steht für *Application Programming Interface* und bezeichnet eine Programmierschnittstelle, die Soft- und Hardwarekomponenten wie Anwendungen, Festplatten oder Benutzeroberflächen verbindet. Sie vereinheitlicht die Datenübergabe zwischen Programmteilen, etwa Modulen, und Programmen.

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 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.

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 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.

OCR OCR steht für *Optical Character Recognition* und bezeichnet die automatische Texterkennung von gedruckten Texten, d. h. ein Computer „liest“ ein eingescanntes Dokument, erkennt und erfasst den Text darin und generiert daraufhin eine elektronische Version.

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.

PDF PDF steht für *Portable Document Format* . Es handelt sich um ein plattformunabhängiges Dateiformat, dessen Inhalt auf jedem Gerät und in jedem Programm originalgetreu wiedergegeben wird. PDF-Dateien können Bilddateien (z. B. Scans von Texten) oder computerlesbarer Text sein. Ein lesbares PDF ist entweder ein OCRter Scan oder ein am Computer erstellter Text.

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.

Query *Query* bedeutet „Abfrage“ oder „Frage“ und bezeichnet eine computergestützte Abfrage zur Analyse eines Textes. Um Datenbestände zu durchsuchen, werden Abfragesprachen eingesetzt, die *Queries* (Anfragen) an den Datenbestand senden. So bilden alle möglichen Queries zusammen die *Query Language* eines Tools.

Reguläre Ausdrücke Reguläre Ausdrücke, auch *Regular Expressions* oder *RegEx* genannt, sind standardisierte Zeichenketten zur Beschreibung von Mengen von Zeichenketten mit Hilfe bestimmter syntaktischer Regeln, die in Abfrage- und Programmiersprachen (z. B. in Wort, CATMA, Python, R usw.) für unterschiedliche Problemlösungen verwendet werden. Sie können beispielsweise als Filterkriterien in der Textsuche oder in Texteditoren (z. B. in Word oder OpenOffice) zum „Suchen und Ersetzen“ von bestimmten Begriffen genutzt werden.

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.

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.

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.