Toolbeitrag: SentText

Marie Flüh [1]

1. Universität Hamburg

Thema:	Sentimentanalyse	DOI:	10.48694/fortext.3799
Jahrgang:	1	Ausgabe:	7
Erscheinungsdatum:	2024-07-10	Erstveröffentlichung:	2020-06-01 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.

Der Workflow bei SentText: Upload: Hochladen einer oder mehrerer TXT- oder XML-Dateien, Auswahl eines Sentimentwörterbuchs oder Upload eines eigenen Sentimentwörterbuchs, individuelle Anpassung der Analyseparameter und Start der Analyse; vierteiliges Interface v. r. n. l.: Text mit farblich markierten Sentimentwörtern, Visualisierungen der Analyseergebnisse, Organisation der Textdateien, Infopanel mit Analyseparamtern und Menü; Output: Download der Visualisierungen als CSV-, PNG- oder XML-Datei

- **Systemanforderungen:** Das Tool ist webbasiert und am besten über Google Chrome oder Firefox nutzbar
- **Stand der Entwicklung:** Entwicklung der Testdemo 2019 und 2020; Überarbeitung und Verlagerung der Endversion auf die Server der Universität Regensburg 2020
- **Herausgeber:** Universität Regensburg: Johanna Dangel (Entwicklung) und Thomas Schmidt
- **Lizenz:** Kostenfrei (Open Source)
- **Weblink:** http://thomasschmidtur.pythonanywhere.com/
- **Im- und Export:** Import von Dateien im TXT (vgl. Reintext-Version)- und XML-Format; Export aller Visualisierungen im CSV-, PNG- oder XML-Format
- **Sprachen:** Deutsch

1. Für welche Fragestellungen kann SentText eingesetzt werden?

Mit SentText können Sie deutschsprachige literarische Texte aller Epochen hinsichtlich der hierin enthaltenen positiven oder negativen Haltungen analysieren lassen. Das Tool richtet sich ausdrücklich an Nutzer*innen mit literaturwissenschaftlichen Forschungsinteressen. Im Rahmen der Entwicklung wurden entsprechende Bedarfsanalysen durchgeführt und in die Fortentwicklung integriert. Das Tool besticht u. a. durch das intuitiv bedienbare Interface (vgl. GUI), das Nutzer*innen bar methodischer Vor- oder Programmierkenntnisse einen Einstieg in die lexikonbasierte Sentimentanalyse (Flüh 2024) ermöglicht. Es eignet sich, um die Polarität (positiv, negativ, neutral) literarischer Texte zu untersuchen und darauf aufbauend Aussagen über die in einem Text vorherrschende positive oder negative Stimmung treffen zu können. Daran anschließend ließe sich nach emotionstragenden Textstrukturen fragen.
Forschungsfragen, die sich bearbeiten lassen, sind z. B.: Welche Sentiment-tragenden Wörter (vgl. SBW) finden sich in Robert Musils Roman *Der Mann ohne Eigenschaften* und herrscht eine überwiegend positive oder negative Stimmung? Verweisen Sentiment-tragende Wörter in Franz Kafkas *Die Verwandlung* auf bestimmte emotionale Zustände und in welcher textuellen Gestalt erscheinen Emotionen in Kafkas Erzählung? Im Rahmen einer Korpusanalyse könnten Sie bspw. untersuchen, mit welcher Terminologie Gefühle in unterschiedlichen Epochen oder literarischen Gattungen zum Ausdruck gebracht wurden.

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

Funktionen:

- Import und lexikonbasierte Sentimentanalyse von Einzeltexten oder Textkorpora (vgl. Korpus)
- Auswahl aus zwei Sentimentwörterbüchern (SentiWS oder BAWL-R, darauf basiert die Berechnung der Sentimentwerte (vgl. Sentimentwert))
- Import eines zuvor selbst erstellten Sentimentwörterbuchs (im CSV-Format) als individuelle Analysegrundlage, die auf das (historische) Vokabular der Textgrundlage ausgerichtet ist
- Feinjustierung der Analyseparameter durch An- oder Ausschalten der folgenden Kenngrößen: Lemmatisierung (vgl. Lemmatisieren), Negationen, Case Sensitivity, Stoppwortliste
- Manuelle Erweiterung der Stoppwortliste
- Analyseergebnisse: Berechnung des durchschnittlichen Sentimentwerts des Textes, Anzeige aller positiven (rot), negativen (blau) oder neutralen (gelb) Sentiment-tragenden Wörter (vgl. SBW) und deren Sentimentwerten im gesamten Dokument
- Spezifische Informationen zum Sentimentwert eines markierten Textabschnitts
- Manuelle Korrektur: Markieren eines Wortes im Textpanel und Vergabe eines Sentimentwerts ermöglichen die manuelle Korrektur falsch erkannter bzw. die Ergänzung nicht erkannter Sentimentwörter
- Visualisierungen der Ergebnisse: Diagramme zur Gewichtung der Polarität des gesamten Textes oder Textkorpus (Barchart: absolute Polarität, normalisierte Polarität und Polaritäten der sentiment bearing words (vgl. SBW)), Diagramme auf Wortebene (Kreisdiagramm: Verteilung der negativen und positiven Wörter, Wordcloud der vermehrt vorkommenden positiven und negativen Sentiment-tragenden Wörter und Barchart der acht am häufigsten auftretenden positiven bzw. negativen Wörter), Diagramme auf Satzebene (Kreisdiagramm: Verteilung der Sätze mit negativer bzw. positiver Valenz; interaktiver Zeitstrahl mit Textrückbezug: Entwicklung der Sentimente im Text oder Korpus; Verzeichnis der zehn Sätze mit den höchsten Sentimentwerten)
- Organisation des Textkorpus: Überblick über gesamte Textgrundlage, Organisation aller zu Beginn hochgeladenen Texte per drag and drop zu bspw. autor*innenspezifischen Textkorpora in unterschiedlichen Ordnern, Ausführung einer vergleichenden Sentimentanalyse mehrerer Ordner (= Textkorpora) oder von Einzeltexten
- Export aller Visualisierungen der Analyseergebnisse

Zuverlässigkeit: Das Tool funktioniert zuverlässig. Bei steigender Anzahl und Größe der Textdateien nimmt der Analyseprozess mehr Zeit in Anspruch. Gleiches gilt für die Auswahl weiterer Analyseparameter wie bspw. die Lemmatisierung. Die manuelle Korrekturmöglichkeit stellt ein wichtiges Feature dar. Wird bspw. in dem Satz *„Nun, die Hoffnung ist noch nicht gänzlich aufgegeben; habe ich einmal das Geld beisammen, um die Schuld der Eltern an ihn abzuzahlen – es dürfte noch fünf bis sechs Jahre dauern –, mache ich die Sache unbedingt.“* „Hoffnung“ aufgrund der Negation „nicht“ ein negativer Sentimentwert zugewiesen, können Sie dies korrigieren – Schließlich lässt sich Hoffnung in diesem Fall durchaus als positive Empfindung interpretieren. Gleichzeitig entsteht ein didaktischer Mehrwert, da die Arbeitsweise lexikonbasierter Sentimentanalysen deutlich wird.

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

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

Ein separates Handbuch ist nicht erhältlich, stellt aber auch keine Notwendigkeit dar. Hilfreiche Informationen für Erstnutzer*innen über Sentimentanalysen finden Sie unter „About“. Fortgeschrittene Nutzer*innen können durch die Auswahl supplementärer Analyseparameter die Standardeinstellungen verfeinern („More options (advanced user)“-Button), die Funktionen werden jeweils anhand kurzer Beispiele veranschaulicht. Beschreibungen und Anwendungshinweise zu einzelnen Elementen der grafischen Benutzeroberflächer erscheinen beim Hovern über das Info-Symbol. Sämtliche Analyseparamter können Sie unter „Documentation“ einsehen. Möglichkeiten zur Kontaktaufnahme und einer persönlichen Nutzerbetreuung finden Sie unter der Rubrik

„Contact“.

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

Da es sich bei SentText um eine Neuerscheinung handelt, konnte der literaturwissenschaftliche Mehrwert des Tools bisher noch nicht nachgewiesen werden. Obwohl derzeit keine wissenschaftlichen Artikel aus interpretativen literaturwissenschaftlichen Forschungseinrichtungen existieren, die das Tool nennen, finden literaturwissenschaftlich ausgerichtete lexikonbasierte Sentimentanalysen durchaus Anwendung (Schmidt und Burghardt 2018; Nalisnick und Baird 2013; Mohammad 2013). Für Unterstützung beim Einsatz von SentText in Forschung oder Lehre stehen die Herausgeber zur Verfügung.

5. Unterstützt SentText kollaboratives Arbeiten?

Nein, es kann nicht kollaborativ gearbeitet werden.

6. Sind meine Daten bei SentText sicher?

Ja, das Tool läuft momentan über den Server-Anbieter pythonanywhere. Im Verlauf des Jahres ist eine Übertragung auf die Informatik-Server der Universität Regensburg geplant. Für die Verwendung müssen Sie keine personenbezogenen Daten angegeben. Sobald Sie eine Sitzung schließen oder per Klick auf „NEW SENTIMENT ANALYSIS“ eine weitere Analyse vornehmen, gehen Ihre Analysedaten verloren.

Externe und weiterführende Links

- BAWL-R: https://web.archive.org/save/https://www.ewi-psy.fu-berlin.de/psychologie/arbeitsbereiche/allgpsy/Download/BAWL-R/index.html (Letzter Zugriff: 16.09.2024)
- Pythonanywhere: https://web.archive.org/save/https://www.pythonanywhere.com/ (Letzter Zugriff: 28.07.2024)
- SentiWS: https://web.archive.org/save/https://wortschatz.uni-leipzig.de/de/download> (Letzter Zugriff: 28.07.2024)

Bibliographie

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.

Mohammad, Saif. 2013. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. https://arxiv.org/pdf/1309.5909.pdf.

Nalisnick, Eric T. und Henry S. Baird. 2013. Character-to-Character Sentiment Analysis in Shakespeare's Plays. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 479–483. Sofia, Bulgaria: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics.

Schmidt, Thomas und Manuel Burghardt. 2018. An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing. In: *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 139–149. Santa Fe, New Mexico: Association for Computational Linguistics.

Schmidt, Thomas, Manuel Burghardt und Katrin Dennerlein. 2018. „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.

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/infdh2018-16,.

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.

Case Sensitivity Unter Case Sensitivity versteht man in der Regel die Berücksichtigung von Groß- und Kleinschreibung von Textelementen bei der Datenverarbeitung. Diese ist unter anderem für die Einstellung von Such- und Analysekriterien bei Tools für die digitale Textanalyse von Bedeutung.

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.

Feature Unter Features können Einzelfunktionen eines Tools verstanden werden, die beispielsweise komplexe Funktionen wie die Visualisierung eines Textes als Wordcloud ermöglichen, oder auch kleinere Funktionseinheiten wie den Abgleich einzelner Spracheigenschaften (Properties) mit annotierten Beispieltexten darstellen.

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.

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 (Textauszeichung) Die Textauszeichnung (eng. *Markup*) fällt in den Bereich der Daten- bzw. Textverarbeitung, genauer in das Gebiet der Textformatierung, welche durch Auszeichnungssprachen wie XML implementiert wird. Dabei geht es um die Beschreibung, wie einzelne Elemente eines Textes beispielsweise auf Webseiten grafisch dargestellt werden sollen.

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.

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.

Property Property steht für „Eigenschaft“, „Komponente“ oder „Attribut“. In der automatischen Annotation dienen konkrete Worteigenschaften wie Groß- und Kleinschreibung zur Klassifizierung von Wörtern oder Phrasen. Durch die Berücksichtigung solcher Eigenschaften in den Features eines Tools kann maschinelles Lernen bestimmter Phänomene umgesetzt werden. In der manuellen Annotation können als Properties auch Eigenschaften von Annotationen benannt 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.

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.

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.

TEI Die *Text Encoding Initiative* (TEI) ist ein Konsortium, das gemeinsam einen Standard für die Darstellung von Texten in digitaler Form entwickelt. Die TEI bietet beispielsweise Standards zur Kodierung von gedruckten Werken und zur Auszeichnung von sprachlichen Informationen in maschinenlesbaren Texten (siehe auch XML und Markup).

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.

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.

XML XML steht für *Extensible Markup Language* und ist eine Form von Markup Language, die sowohl computer- als auch menschenlesbar und hochgradig anpassbar ist. Dabei werden Textdateien hierarchisch strukturiert dargestellt und Zusatzinformationen i. d. R. in einer anderen Farbe als der eigentliche (schwarz gedruckte) Text dargestellt. Eine standardisierte Form von XML ist das TEI-XML.