Warum wird was wie klassifiziert? Scalable Reading + Explainable AI am Beispiel historischer Lebensverläufe

Brookshire, Patrick Daniel
Patrick.Brookshire@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz / Universität zu Köln, Deutschland
ORCID: 0000-0002-5843-7577

Einleitung

Eine Klassifikation von Textabschnitten ist im DH-Kontext häufig „under construction“, da vorliegende Verfahren viele Varianten aufweisen, aber nicht domänen-spezifisch genug sind. Deshalb ist stets eine Evaluation mit dem konkreten Datensatz nötig. Zudem sind bei besonders kontextabhängigen Tasks, für die Sentimentanalysen ein verbreitetes Beispiel sind, Deep-Learning-Methoden performanter, dafür aber weniger interpretierbar als bspw. Lexikon-basierte Verfahren (Singh und Singh, 2021; Schmidt et al., 2022; Rebora et al., 2023). Eine mögliche Lösung sind Explainability-Modelle wie *LIME* (Ribeiro et al., 2016), *SHAP* (Lundberg und Lee, 2017) oder *Transformer-Explainability* (Chefer et al., 2021). Allerdings erfordern diese mitunter mehr Rechenleistung als das Finetuning eines Klassifikationsmodells (Brookshire und Reiter, 2024, 99f.). Nichtsdestotrotz lassen sie sich in einen an das Scalable Reading (vgl. Weitin 2017) angelehnten Workflow einbinden, mit dem Analysen iterativ in ihrer Validität verbessert werden können. Dieser wird im folgenden am Beispiel der Untersuchung von Polaritätsverläufen illustriert, welche sich bei biographischen Daten anbietet (Dennis-Henderson et al., 2020, 96; Koncar et al., 2020, 8–10; Faull und McGuire, 2022, 143f.).

Biographiekorpus

In einer Pilotstudie werden Daten des retrodigitalisierten biographischen Nachschlagewerks *Allgemeine Deutsche Biographie* (Reinert et al., 2015) nachgenutzt. Es umfasst insgesamt über 25.000 Artikel zu Personen, die vor 1900 geboren sind. Davon werden hier knapp 3.600 Personen, die zwischen 1750 und 1850 geboren und gestorben sind, berücksichtigt. Diese Zahl wurde durch ein teil-randomisiertes Subsampling auf 360 reduziert, welches sicherstellt, dass jeder Text mind. 10 Sätze umfasst. Desweiteren erfolgte als einziger Preprocessing-Schritt eine Satztokenisierung mit *stanza* (Qi et al., 2020).

Methodologie und inhaltliche Erkenntnisse

Aus technischer Sicht umfasst das vorgestellte Verfahren die fünf Module *Klassifikation*, *Ordinal-Transformation*, *Segmentierung*, *Aggregierung* und *Erklärung*, wobei die ersten vier einem Distant-Reading-Ansatz folgen. So lassen sich etwa Polaritätsverläufe der zu untersuchenden historischen Biographien folgendermaßen operationalisieren: Zunächst weist ein passendes **Klassifikations**-Modell (Brookshire und Reiter, 2024) den Sätzen automatisch Sentiment-Labels zu. Anschließend werden diese nominalen Kategorien durch ein Mapping (negativ: -1, neutral: 0, positiv: 1) in ordinale (= Polarität) **transformiert**. Im dritten Schritt erfolgt eine **Segmentierung** in bspw. gleichlange Textabschnitte (Steps), um Texte unterschiedlicher Länge miteinander vergleichen zu können. Abschließend werden die Werte pro Segment **aggregiert** – üblicherweise mit Mittelwerten, was letztlich eine Parametrisierung ist –, um einen Gesamt-Verlauf der 360 Biographien zu erzeugen:

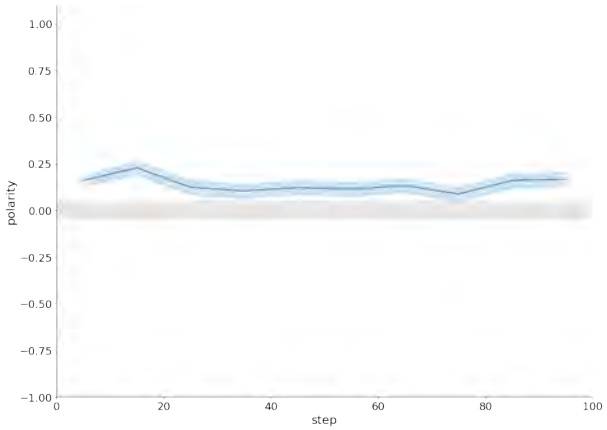


Abb. 1: Durchschnittlicher Polaritätsverlauf mit Hervorhebung des neutralen Bereichs

Durch die Markierung des neutralen Bereichs mit gängigen Schwellenwerten (± 0,05; vgl. Hutto und Gilbert, 2014, 224) wird deutlich, dass das Korpus einen leicht positiven Trend mit nur wenigen Schwankungen aufweist. Allerdings zeichnet ein Reinzoomen im Sinne des Scalable Readings auf 36 zufällig ausgewählte Einzelbiographien ein deutlich anderes Bild:

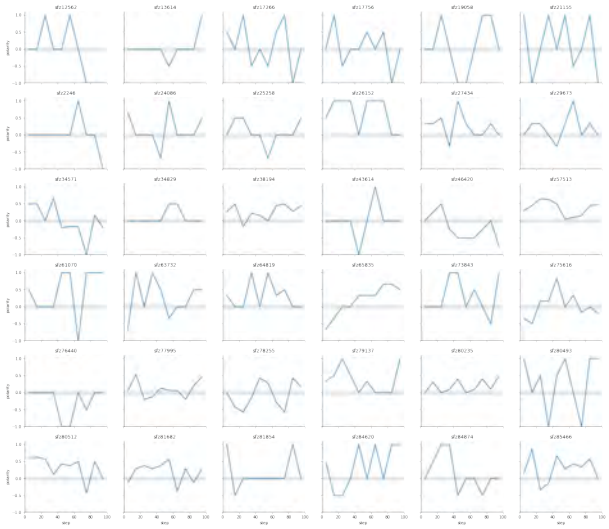
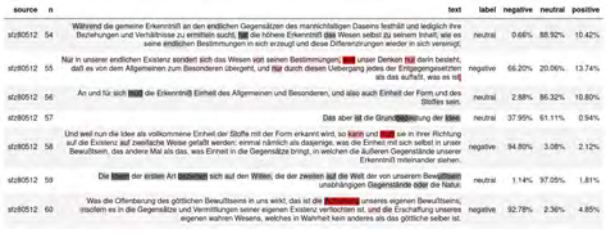


Abb. 2: 36 Einzel-Polaritätsverläufe mit Hervorhebung des neutralen Bereichs

Die Variabilität ist bei einem individuellen Genre wie Biographien nicht überraschend, aber bei der Analyse größerer Datenmengen wegen der Tendenz von Durchnittswerten zum neutralen Bereich (vgl. Jockers, 2015) nur durch Skalenwechsel zu beobachten.

Aufgrund der nicht perfekten automatischen Annotationen kann auch diese Einzelansicht mitunter täuschen, weshalb im vorgelegten Verfahren eine am Close-Reading orientierte **Explainability**-Schleife als fünftes Modul vorgesehen ist. In dieser kann bspw. der starke Abfall in den negativen Bereich bei 75% der Biographie der Person mit der ID *sfz80512* erklärt werden.



source	n	text	label	negative	neutral	positive
sfz80512	54	Während die gemeine Erkenntniß an den endlichen Gegensätzen des mannichfaltigen Daseins festhält und lediglich ihre Beziehungen und Verhältnisse zu ermitteln sucht, hat die höhere Erkenntniß das Wesen selbst zu seinem Inhalt, wie es seine endlichen Bestimmungen in sich erzeugt und diese Differenzirungen wieder in sich vereinigt;	neutral	0.66%	88.92%	10.42%
sfz80512	55	Nur in unserer endlichen Existenz sondert sich das Wesen von seinen Bestimmungen; nur unser Denken hat darin besteht, daß es von dem Allgemeinen zum Besonderen übergeht, und nur durch diesen Uebergang jedes der Entgegengesetzten als das aufhält, was es ist;	negative	66.20%	20.06%	13.74%
sfz80512	56	An und für sich hat die Erkenntniß Einheit des Allgemeinen und Besonderen, und also auch Einheit der Form und des Stoffes sein.	neutral	2.88%	96.32%	10.80%
sfz80512	57	Das aber ist die Grundbestimmung der Idee.	neutral	37.95%	61.11%	0.94%
sfz80512	58	Und weil nun die Idee als vollkommene Einheit der Stoffe mit der Form erkannt wird, so kann und muß sie in ihrer Richtung auf die Existenz auf zweifache Weise gefaßt werden: einmal nämlich als dasjenige, was die Einheit mit sich selbst in unser Bewußtsein, das andere Mal als das, was Einheit in die Gegenstände bringt, in welchen die äußeren Gegenstände unserer Erkenntniß miteinander stehen.	negative	94.80%	3.08%	2.12%
sfz80512	59	Die Idee der ersten Art besteht sich auf den Willen, die der zweiten auf die Welt der von unserem Bewußtsein unabhängigen Gegenstände oder die Natur.	neutral	1.14%	97.05%	1.81%
sfz80512	60	Was die Offenbarung des göttlichen Bewußtseins in uns wirkt, das ist die Aufhebung unseres eigenen Bewußtseins, insofern es in die Gegensätze und Vermittlungen seiner eigenen Existenz verflochten ist, und die Erschaffung unseres eigenen wahren Wesens, welches in Wahrheit kein anderes als das göttliche selber ist.	negative	92.78%	2.36%	4.85%

Abb. 3: Einfärbung der Auswirkung einzelner Subwordtokens auf die Klassifikation

So gehen die hier durch Farbsättigung visualisierten Ergebnisse des (aus Performance-Gründen gewählten) Explainability-Modells von Chefer et al. (2021) über die reinen Klassifikationswahrscheinlichkeiten hinaus. Denn sie illustrieren, dass die negativ-Labels auf Subword-Tokens wie „nur“, „muß“ und „Aufhebung“ zurückgehen, aber den gegebenen neutral formulierten philosophischen Abschnitt nur bedingt abbilden. In diesem Fall wäre also eine manuelle Bereinigung der entsprechenden Labels sinnvoll und dank der mitvisualisierten Position im Gesamtdatensatz durch Quellen-ID und Satznummer auch möglich. Zudem ist auch ein erneutes Finetuning mit diesen Datensätzen denkbar.

Fazit und Ausblick

Das vorgestellte Verfahren zeigt, wie sich Fehlklassifikationen, die für nachgelagerte Analyseschritte besonders relevant sind, durch eine Kombination aus Scalable-Reading- und Explainability-Ansätzen gezielt identifizieren und ggf. manuell korrigieren lassen. Denn zur Steigerung der Validität ist es durch Aggregierungsschritte nicht immer nötig, alle Datensätze zu bereinigen – und je nach personeller und technischer Ausstattung auch nicht immer umsetzbar. Auch wenn das Verfahren hier am Beispiel von Sentimentwerten illustriert wurde, ist es grundsätzlich auf beliebige Klassifikationsaufgaben anwendbar. Daher ist derzeit ein entsprechender Ausbau in Richtung nominal-skalierter Kategorien „under construction“. Zudem ist angedacht, die Modularisierung so konsequent umzusetzen, dass künftig neben beliebigen Klassifikationsmodellen auch bei den übrigen Modulen beliebige Komponenten angedockt werden können, um dem Variantenreichtum gerecht zu werden.

Bibliographie

Brookshire, Patrick D. und Nils Reiter. 2024. „Modeling Moravian Memoirs: Ternary Sentiment Analysis in a Low Resource Setting“. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 91–100. https://aclanthology.org/2024.latechclfl-1.10.

Chefer, Hila, Shir Gur und Lior Wolf. 2021. „Transformer Interpretability Beyond Attention Visualization“. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 782–791. https://doi.org/10.1109/CVPR46437.2021.00084.

Dennis-Henderson, Ashley, Matthew Roughan, Lewis Mitchell und Jonathan Tuke. 2020. „Life Still Goes on: Analysing Australian WW1 Diaries through Distant Reading“. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 90–104. https://aclanthology.org/2020.latechclfl-1.11/.

Faull, Katherine und Michael McGuire. 2022. „Analyzing Moravian Feelings Using Computational Methods to Ask Questions about Norms and Sentiments in Eighteenth-Century Moravian Lebensläufe“. *Journal of Moravian History* 22 (2): 125–149. https://doi.org/10.5325/jmorahist.22.2.0125.

Hutto, Clayton J. und Eric Gilbert. 2014. „VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text“. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 216–222. https://doi.org/10.1609/icwsm.v8i1.14550.

Jockers, Matthew L. 2015. „Revealing Sentiment and Plot Arcs with the Syuzhet

Package“ https://www.matthewjockers.net/2015/02/02/syuzhet/ (zugegriffen: 22.07.2024).

Koncar, Philipp, Alexandra Fuchs, Elisabeth Hobisch, Bernhard C. Geiger, Martina Scholger und Denis Helic. 2020. „Text Sentiment in the Age of Enlightenment: An Analysis of Spectator Periodicals“. *Applied Network Science* 5 (1): 1–32. https://doi.org/10.1007/s41109-020-00269-z.

Lundberg, Scott M. und Su-In Lee. 2017. „A Unified Approach to Interpreting Model Predictions“. *Advances in Neural Information Processing Systems* 30. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton und Christopher D. Manning. 2020. „Stanza: A Python Natural Language Processing Toolkit for Many Human Languages“. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14.

Rebora, Simone, Marina Lehmann, Anne Heumann, Wei Ding und Gerhard Lauer. 2023. „Comparing ChatGPT to Human Raters and Sentiment Analysis Tools for German Children’s Literature“. In *Proceedings of the Computational Humanities Research Conference*, 333–343. https://ceur-ws.org/Vol-3558/paper3340.pdf.

Reinert, Matthias, Maximilian Schrott und Bernhard Ebneth. 2015. „From Biographies to Data Curation – the Making of www.deutsche-biographie.de“. In *Proceedings of the First Conference on Biographical Data in a Digital World*, 13–19. https://ceur-ws.org/Vol-1399/paper3.pdf.

Ribeiro, Marco Tulio, Sameer Singh und Carlos Guestrin. 2016. „‚Why Should I Trust You?‘: Explaining the Predictions of Any Classifier“. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778.

Schmidt, Thomas, Katrin Dennerlein und Christian Wolff. 2022. „Evaluation computergestützter Verfahren der Emotionsklassifikation für deutschsprachige Dramen um 1800“. In *8. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e. V*. https://doi.org/10.5281/zenodo.6328169.

Singh, Loitongbam Gyanendro und Sanasam Ranbir Singh. 2021. „Empirical Study of Sentiment Analysis Tools and Techniques on Societal Topics“. *Journal of Intelligent Information Systems* 56 (2): 379–407. https://doi.org/10.1007/s10844-020-00616-7.

Weitin, Thomas. 2017. „Scalable Reading“. *Zeitschrift für Literaturwissenschaft und Linguistik* 47 (1): 1–6. https://doi.org/10.1007/s41244-017-0048-4.