

Hinterlistig – schelmisch – treulos – Sentiment Analyse in Texten des 19. Jahrhunderts: Eine exemplarische Analyse für Länder und Ethnien

Wodausch, David

wodausch@uni-hildesheim.de
Universität Hildesheim, Deutschland

Fiedler, Maik

fiedler@gei.de
Georg-Eckert-Institut – Leibniz-Institut für internationale
Schulbuchforschung (GEI)

Heuwing, Ben

heuwing@uni-hildesheim.de
Universität Hildesheim, Deutschland

Mandl, Thomas

mandl@uni-hildesheim.de
Universität Hildesheim, Deutschland

Abstract

Sentiment Analyse ist eine Standard-Methode des Text Mining und prüft, was in Texten meinungsbehaftet behandelt wird. Damit ist es auch für Digital Humanities sehr interessant. Viele Ansätze basieren auf einfachen Listen, sogenannten Sentiment Lexika, welche sich aber nicht vollständig auf andere Domänen und auch kaum auf andere Zeiträume übertragen lassen. Dieser Beitrag zeigt beispielhaft, wie eine solche Liste methodisch für Schulbücher des 19. und frühen 20. Jhds. erstellt werden kann. Am Beispiel der Behandlung des Judentums in den historischen Schulbüchern wird gezeigt, dass automatische Ansätze des Text Mining alleine nicht ausreichen, um die Komplexität der Bewertung zu erfassen. Erst durch die Verknüpfung intellektueller und digitaler Ansätze entstehen interessante Ergebnisse. Zudem wird gezeigt, dass ein aktuelles Sentiment Lexikon für diese historischen Epoche ungeeignet ist.

Einleitung

Text Mining ist in den Digital Humanities eine häufig eingesetzte Technologie mit vielen Facetten. Die Sentiment Analyse versucht zu analysieren, welche Meinungen ein Text zu einer Entität ausdrückt. Dazu werden die im Kontext von Erwähnungen der Entität vorkommenden Stellen ausgewertet (Liu 2012). Häufig ist eine Meinung nicht leicht maschinell zu erkennen, so dass leichtgewichtige linguistische Ansätze (Struß 2016) oder auch tiefgehende Analysen notwendig sind (Somasundaran et al. 2008). Für viele Anwendungen werden aber einfache listenbasierte Ansätze genutzt, welche das Vorkommen von positiven und negativen Wörtern einer Sprache im Umfeld einer Entität oder einen Konzept, das mehrere Wörter berücksichtigt, auswerten.

In der hier geschilderten Untersuchung sollte geprüft werden, ob eine Liste für historische Texte effizient erstellt werden könnte und inwieweit damit fachwissenschaftliche Fragestellungen analysiert werden könnten. Im Rahmen des Projektes „Welt der Kinder“ (De Luca 2014) stand eine Kollektion von über 3500 Schulbüchern des 19. und frühen 20. Jhd. ebenso zur Verfügung wie Werkzeuge zur Analyse (wdk.gei.de). Die Basis für die Systementwicklung bildete eine Analyse der Informationsbedürfnisse der von Historikern (Heuwing et al. 2016). Unter anderem wurden Topic Models und darauf basierende Filter entwickelt (Schober & Gurevych 2015).

Schulbücher eignen sich besonders als Gegenstand der Forschung zum 19. Jhd. Diese Bücher waren oft die einzigen Medien, welche Kinder ausgesetzt werden. Sie prägten somit ihr Weltbild nachhaltig (Fuchs 2014). Die Dokumente sind durch OCR in Volltexte umgewandelt. Dabei ergibt sich insbesondere für Fraktur eine Fehlerrate. Diese ist jedoch gering und eine quantitative Auswertung ist gleichwohl möglich.

Methode

Ziel des Vorgehens war es, ein erstes und vorläufiges Sentiment Wörterbuch aus Schulbüchern des 19. Jhds. zu erstellen und zu prüfen, inwieweit dies sich von einer aktuellen Liste unterscheidet. Dazu wurde folgendes methodische Vorgehen gewählt, das zwar nicht ausreicht, ein vollständiges Sentiment Lexikon für die Kollektion zu erstellen, aber doch zu interessanten Resultaten führt. Durch die Auswahl von Begriffen, die häufig mit Sentiment verbunden sind, sollten Textstellen identifiziert werden, die dann einer manuellen Auswertung unterzogen wurden. Bei dieser Bewertung wurden die meinungstragenden Phrasen als positiv oder negativ bewertet.

Konkret wurde zunächst aus dem Gesamtkorpus drei Untermengen ausgewählt, wobei Anfragen gewählt wurden, die nach ersten Hypothesen aus Perspektive der Geschichtswissenschaft zu Aussagen mit Meinungen führen würden.

- In einer Untermenge, welche zu Völkern und Staaten generiert wurde, erfolgte eine Recherche nach den Suchbegriffen Deutschland und Frankreich.
- Der Name Napoleon Bonaparte wurde in Jungen und Mädchenbüchern getrennt recherchiert, um so mögliche unterschiedliche Darstellungen zu erkennen.
- Mit dem Wortfeld Jude wurde in Büchern aus zwei Zeiträumen gesucht. Diese umfassten 1850-1859 und 1910-1919. Dabei stand die Hypothese im Raum, dass die verwendete Sprache des Zeitraums 1850 bis 1859 einen weniger radikalisierten Antisemitismus aufweist als die Analyse der Textquellen von 1910 bis 1919. Diese Einschätzung ist auf den ab 1879 bis nach den Ersten Weltkrieg sich radikalisierenden Antisemitismus zurückzuführen.

Das Werkzeug AntConc erlaubt es, Konkordanzen zu extrahieren. Die Suchbegriffe wurden als zentrale Wörter dargestellt und AntConc liefert dann den Kontext der Phrase. Die Sortierung der Phrasen erfolgt nach dem Maß Mutual Information.

So wurden zu jeder Suche und damit zu jedem der sechs Sub-Korpora jeweils ca. 100 Phrasen manuell ausgewertet und klassifiziert. Dabei wurde festgehalten, ob eine meinungstragende Phrase vorliegt. Zudem wurde über die Polarität entschieden, also ob die Meinung positiv oder negativ ist. Diese manuelle Annotation semantischer Werte in Textabschnitten erfolgte, um darauf aufbauend das Sentiment Lexikon zu erstellen.

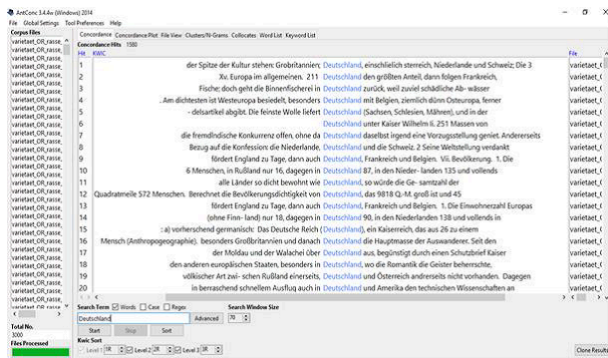


Abbildung 1: Kollokationen in AntConc

Ergebnisse

Für die erste Textmenge ergaben sich dabei 270 unterschiedliche Kollokationen, die eine positive oder negative interpretierte Meinung zu Deutschland oder Frankreich besitzen. Dabei wird Frankreich negativer dargestellt, 49% der meinungstragenden Phrasen sind negativ. Bei Deutschland sind es 39%. Diese beziehen sich beispielsweise auf die Situation vor 1871 und kennzeichnen diese als Zersplitterung.

Für die Textstellen zu Napoleon Bonaparte konnten 174 wertende Phrasen erkannt werden. Allerdings ergaben sich

keine nennenswerten Unterschiede in der Darstellung der Person in Jungen und Mädchenschulbüchern.

Für die qualitative Analyse der beiden Stichproben aus den Subkorpora für das Judentum wurden für den Veröffentlichungszeitraum von 1850 bis 1859 insgesamt 720 und beim zweiten Subkorpora von 1910 bis 1919 insgesamt 825 Konkordanzen untersucht. Dabei wurden insgesamt 231 Kollokationen aus 155 wertenden Konkordanzen gefunden. Insgesamt waren die meisten Vorkommen von Begriffen aus dem Wortfeld Jude und Judentum also nicht meinungsbehaftet.

Die Auswertung ergab, dass durchschnittlich 76,5% aller wertenden Kollokationen in den untersuchten Konkordanzen der Subkorpora als negativ klassifiziert wurden (Subkorpora (1): 102 negativ (77%); Subkorpora (2): 75 negativ (76%)). Somit ist das Verhältnis zwischen negativ wertenden und positiv wertenden Kollokationen beider Sub-Korpora annähernd gleich, wobei negative Äußerungen klar überwiegen. Allerdings ergibt sich aus dieser rein quantitativen Betrachtung keine Meinungsentwicklung gegenüber Juden während des Kaiserreichs, wie sie die Hypothese vermutet hatte.

In der Folge wurden die Kollokationen manuell in sechs, in sich homogene Gruppen geclustert, welche jeweils den lokalen Kontext und das Auftreten einer Kollokation beschreiben: Religion; Synonym für Jude; positive Eigenschaft Jude; negative Eigenschaft Jude; Juden im Gesellschaftsleben und Sonstiges. Entwickelt bzw. festgelegt wurden die sechs Cluster auf Grundlage der Kollokationen.

Die Analyse der Vorkommenshäufigkeit der sechs Cluster ergab eine zeitliche Veränderung. Im Zusammenhang mit Religion wurde das Judentum im früheren Subkorpora ab 1850 häufiger erwähnt als im Zeitraum von 1910 bis 1919. Im späteren Zeitraum wird das Judentum häufiger im Zusammenhang mit dem gesellschaftlichen Leben erwähnt. Diese spiegelt möglicherweise den Wandel des Judenhasses während der zweiten Hälfte des 19. Jahrhunderts auf Grundlage religiöser Motive zu einer biologisch-rassistischen Begründung.

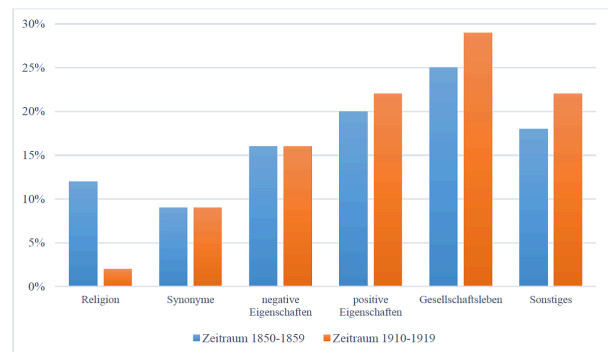


Abbildung 2: Cluster zum Wortfeld Jude/Judentum

Insgesamt wurden 225 einmalig auftretende meinungstragende Begriffe gefunden. Von diesen sind nur 44% in einer aktuellen Liste des Deutschen vorhanden (Sentiment Wortschatz der Universität Leipzig). Zwar ist die verwendete Liste deutlich länger und aus umfangreicheren Studien hervorgegangen, doch allein dieser Vergleich zeigt, dass das historische Vokabular sich deutlich unterscheidet. Nicht enthalten sind beispielsweise: verachtet, zerstreut, unkünstlerisch oder halsstarrig.

Auffällig ist, dass im historischen Korpus deutlich mehr meinungstragende Substantive enthalten sind als im aktuellen Korpus, in dem Adjektive überwiegen. Dies müsste weiter untersucht werden. Möglicherweise wurden Substantive mit spezifischen Zielen wie etwa einer oberflächlichen Objektivierung eingesetzt. Beispiele für Substantive sind: Zarenwahnsinn, Blitzesschnelle, Engherzigkeit oder Schlappe.

Die vorläufigen Ergebnisse zeigen, dass nur eine Verknüpfung des iterativen Einsatzes analoger manueller und digitaler automatischer Methoden verschiedene Perspektiven auf Texte einnehmen kann und zu sinnvollen Ergebnissen führen kann. Dies fordert auch Stephen Ramsay unter dem Schlagwort Algorithmic Criticism (Ramsay 2008).

Mit Criticism wird dabei nicht die Überprüfung hermeneutischer Hypothesen mit algorithmischen Analysen gemeint, sondern die Reflexion, Explikation und Differenzierung von Geisteswissenschaften und Algorithmen sowie die Einbindung multiperspektivischer Zugänge zu Untersuchungsgegenständen (vgl. Bender, 2016, S. 300f.). Möglichkeiten für weitere Forschung liegen in der Verknüpfung mit anderen Quellen für historisches Vokabular.

Bibliographie

Bender, Michael (2016): Forschungsumgebungen in den Digital Humanities: Nutzerbedarf, Wissenstransfer, Textualität. Walter de Gruyter GmbH & Co KG

De Luca, Ernesto William (2014): Welt der Kinder. Georg-Eckert-Institut – Leibniz-Institut für internationale Schulbuchforschung. URL: <http://welt-der-kinder.gei.de>

Fuchs, Eckhardt (2014): Das Schulbuch in der Forschung. Analysen und Empfehlungen für die Bildungspraxis. Georg-Eckert-Institut für internationale Schulbuch-forschung. Band 4. V & R unipress in Göttingen

Heuwing, Ben / Mandl, Thomas / Womser-Hacker, Christa (2016): "Combining contextual interviews and participative design to define requirements for text analysis of historical media". In: Information Research 21 (4) <http://www.informationr.net/ir/21-4/isic/isic1606.html>

Liu, Bing (2012): Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers

Ramsay, Stephen (2008): "Algorithmic Criticism". In: A Companion to Digital Literary Studies, ed. Susan Schreibman and Ray Siemens. Oxford: Blackwell, <http://www.digitalhumanities.org/companionDLS/>

Schnober, Carsten / Iryna Gurevych (2015): "Combining Topic Models for Corpus Exploration: Applying LDA for Complex Corpus Research Tasks in a Digital Humanities Project." In Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications, TM '15. New York, NY, USA: ACM, 2015, S. 11–20.

Somasundaran, Swapna / Josef Ruppenhofer / Janyce Wiebe (2008) "Discourse level opinion relations: An annotation study." In: Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue. Association for Computational Linguistics.

Strauß, Julia Maria (2016): Multilinguales aspektbasiertes Opinion Mining: Entwicklung eines ressourcenarmen Extraktionsverfahrens und Untersuchung von Nutzerperspektiven. Dissertation Universität Hildesheim.