Gernot Heisenberg, TH Köln & Tina Hees, Questback GmbH

Text Mining-Verfahren zur Analyse offener Antworten in Online-Befragungen im Bereich der Markt- und Medienforschung

Hintergrund, Zielgruppe und Anwendungsfelder

Text Mining ist eine interdisziplinäre Forschungsrichtung, jedoch befassen sich nur sehr wenige Arbeiten bislang mit dem Extrahieren, Analysieren und Erkennen von freien Antworten aus Online-Befragungen.

Dabei haben die Methoden zur inhaltlichen Erfassung und automatischen Analyse von Texten aufgrund der stark gestiegenen Verfügbarkeit und des möglichen Zugriffs zu digitalisierten Textdaten in den letzten Jahren stark an Bedeutung gewonnen. Das Finden relevanter Dokumente für bestimmte Suchanfragen im Bereich Information Retrievals kann bereits sehr gut automatisch ausgeführt werden. Die Herausforderung beim Text Mining liegt darin, dass Informationen in natürlichsprachiger und unstrukturierter Form vorliegen.

Um sie verwenden zu können, müssen diese semantisch gedeutet und die sprachlichen Zusammenhänge erst einmal erkannt werden. Je nach Art der Textdaten und des Informationsbedürfnisses entstehen darüber hinaus – je nach Anwendungsfall – weitere spezifische Herausforderungen.

Text Mining spielt neben offenen Befragungsantworten im Bereich der Markt- und Medienforschung auch für viele andere Anwendungsfelder eine große Rolle.

So wird es zum Beispiel sehr stark im Social Media-Bereich eingesetzt, um dortige Beiträge hinsichtlich ihrer Sentiments auszuwerten, im Customer-Relationship-Management Emails und Posts in Beschwerde- und Nicht-Beschwerde-Emails zu klassifizieren oder Produktbewertungen für Marketingzwecke zu analysieren.

Insbesondere bei großen Textmengen kommt man bei manueller Bearbeitung der Textdaten schnell an die Grenzen eines akzeptablen Aufwandes. Text Mining findet aber auch in weniger zu erwartenden Feldern seine Anwendung.

So zum Beispiel zur Nebenwirkungsforschung in der Medizin (Warrer et al. 2012) oder zur Kommunikationsüberwachung bei der Verbrechensbekämpfung (Keyvanpour, Javideh & Ebrahimi 2011) und dabei insbesondere zur Detektion von „Hate-Speech" in sozialen Netzwerken (Ting et al. 2013).

Text Mining-Ansätze und Methoden

Text Mining ist die Offenlegung und Gewinnung von informativem, nicht-trivialem Wissen aus freiem und unstrukturiertem Text. Dies schließt Methoden von Information Retrieval über die Extraktion von Entitäten und Relationen bis zu Text-Klassifikationen und Clustering mit ein (Kao & Poteet 2007). Es kann als eine Erweiterung des Data Mining auf Texte betrachtet werden (Zeng et al. 2012). Häufig werden dabei auch Methoden aus dem spezifischeren Bereich des Natural Language Processing (NLP) genutzt.

> „Das genaue zu verwendende Methodenset hängt dabei immer von der konkreten Fragestellung ab."

Die Auswertung von freien Antwortfeldern bei Befragungen im Bereich der Markt- und Medienforschung verlangt aber in jedem Fall eine Phrasenextraktion und das Auffinden von wiederkehrenden Themenkategorien (Topics). Sind diese noch in ihrer Stimmung zu bewerten (z. B. bei Mitarbeiterbefragungen bzgl. Arbeitsbedingun-

gen), erfolgt in der Regel noch eine Sentimentanalyse. Im Folgenden werden unterschiedliche Ansätze aus den Bereichen der Phrasenextraktion, Sentimentanalyse und Kategorisierung (Topic Modelling) skizziert.

Phrasenextraktion. Die Extraktion von Phrasen kann auf verschiedene Arten durchgeführt werden, wobei sich als besonders gut die Ansätze Part of Speech-Tagging und Chunking (PoS), Stoppwortgrenzen und Kookkurrenzen erwiesen haben.

PoS-Tagging bedeutet in diesem Zusammenhang, dass der zu analysierende Freitext mit Tags versehen wird, in dem jedem Wort entsprechend seiner morphosyntaktischen Rolle (Kasus, feste Präposition usw.) im Satz eine Wortform wie beispielsweise Nomen, Adjektiv oder Artikel zugeordnet wird. Das Erkennen von Kollokationen, d. h. Ausdrücke aus zwei oder mehr Wörtern, die gemeinsam mehr oder einen anderen Sinn als alleine haben und in einer syntaktischen Beziehung miteinander stehen, wird dann als Chunking bezeichnet.

Eigene Untersuchungen zeigen hierbei, dass das PoS-Tagging und Chunking die relevantesten Phrasen aus größeren Textmengen genauer identifizieren konnte, während die Methode der Stoppwortgrenzen auch aus kleineren Textmustern eine größere Menge relevanter Phrasen extrahieren konnte.

Sentimentanalyse. Für die Sentimentanalyse werden in den meisten Fällen Sentimentlexika genutzt, in denen eine große Sammlung von Wörtern bereits entsprechend ihrer vorrangig assoziierten Wertung mit einem positiven, negativen oder neutralen Label oder mit einem Sentimentwert (Sentimentscore) verknüpft ist.

Die Häufigkeit positiv oder negativ verknüpfter Begriffe in einem Text wird dann als Anhaltspunkt für eine Sentimentbewertung des Inhaltes genutzt (Liu 2012). Darüber hinaus können grammatikalische Textstrukturen für eine zusätzliche Einordnung der Aussagen dienen.

Dabei werden beispielsweise Konditionalsätze, sogenannte Sentiment-Shifter, wie etwa Negationen sowie intensivierende oder abschwächende Formen betrachtet. Die Methoden zur Extraktion der wichtigsten Sentimentmerkmale sind sehr ähnlich zu der bereits aufgeführten Problemstellung der Phrasenextraktion.

Da hier insbesondere die Extraktion der häufigsten Nomen, Nomenkombinationen oder auch Nominalphrasen gefragt ist, wird häufig eine gezielte Extraktion mit Hilfe von PoS-Tags eingesetzt. Die Bestimmung des Sentiments findet dann in der Regel Lexika-basiert statt. Allerdings zeigen eigene Untersuchungen, dass diese Bestimmungen wesentlich akkurater werden, wenn sie um syntaxbasierte Regeln erweitert werden.

Die Anpassungen sollten sowohl allgemeine, als auch spezifische Einflussfaktoren abdecken. Grundlage bildet der in Python implementierte Open Source Valence Aware Dictionary and sEntiment Reasoner (VADER) (Hutto & Gilbert 2014). Eigene syntaktische Regeln lassen sich darin einfach erweitern.

Topic Modeling. Beim Topic Modeling werden die Muster kookkurrierender Terme betrachtet, um semantische Zusammenhänge zu modellieren, die sich dann als Themen (engl.: Latent Topics) äußern, die den Dokumenten zugrunde liegen.

Ein häufig gewählter Ansatz, um ein Dokument als eine Menge solcher zugrundeliegenden Topics zu modellieren, ist die Latent Dirichlet Allocation (LDA). Bei dieser Form des Topic Modelings bilden probabilistische Modelle die Ausgangsbasis, um die zugrunde liegende Struktur von zusammenhängenden Themen abzubilden.

Jedes Dokument wird als Zusammensetzung aus Latent Topics gesehen und jedes Latent Topic wiederum als Zusammensetzung aus Termen. Die Terme werden in dem Topic jeweils entsprechend ihrer Bedeutung für dieses Topic gewichtet.

Ein Term kann dabei auch mehreren Topics zugeordnet werden. Eine andere sehr gute Methode des Topic Modeling ist die Non-Negative Matrix Factorization (NMF) (Lee &

Seung 1999). Die NMF ist eine non-probabilistische Methode. Sie arbeitet mit dem Prinzip der Matrix-Dekomposition.

Dabei wird die Dokument-Term-Matrix V als Ausgangsbasis benötigt, welche in die beiden Submatrizen W und H zerlegt wird, sodass gilt:

$$V \approx WH$$

W kann dabei als Topic-Term-Matrix interpretiert werden und H als Dokument-Topic-Matrix. Aus dieser zerlegten Repräsentationsform kann, wie beim LDA-Algorithmus, ein Topic durch die entsprechenden Top N der höchstgewichteten Terme aus der Matrix und ein Dokument durch seine Zusammensetzung aus Topics dargestellt werden.

Insbesondere die NMF zeigte bei eigenen Untersuchungen eine stärkere Konvergenz und höhere Überschneidungen mit den Topics, die zuvor manuell erstellt wurden. Auch bezogen auf die Interpretierbarkeit (die Zuordnung von Labels für die gefundenen Topics) wurde eine bessere Leistung im Vergleich zur LDA erzielt.

Quellen

Hutto, C. J. & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the 8th International Conference on Weblogs and Social Media (S. 216–225). Palo Alto, CA, USA: AAAI Press.

Kao, A. & Poteet, S. R. (2007). Overview. In A. Kao & S. R. Poteet (Hrsg.), Natural language processing and text mining (S. 1–7). London: Springer Verlag Limited.

Keyvanpour, M. R., Javideh, M. & Ebrahimi, M. R. (2011). Detecting and investigating crime by means of data mining: A general crime matching framework. Procedia Computer Science, 3, 872–880.

Lee, D. D. & Seung, S. (1999). Learning the parts of objects by Non-negative Matrix Factorization. Nature, 401 (6755), 788–791.

Liu, B. (2012). Sentiment analysis and opinion mining. San Rafael, CA, USA: Morgan & Claypool Publishers.

Ting, I.-H., Wang, S.-L., Chi, H.-M. & Wu, J.-S. (2013). Content matters: A study of hate groups detection based on social networks analysis and web mining. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (S. 1196–1201). New York, NY, USA: ACM.

Warrer, P., Hansen, E. H., Juhl-Jensen, L. & Aagaard, L. (2012). Using text-mining techniques in electronic patient records to identify ADRs from medicine use. British Journal of Clinical Pharmacology, 73 (5), 674–684.

Zeng, L., Li, L., Duan, L., Lu, K.,Shi, Z., Wang, M., Wu, W. et al. (2012). Distributed data mining: A survey. Information Technology & Management, 13 (4), 403–409.

Über den/die Autor*in

Technology Arts Sciences TH Köln

Prof. Dr. Gernot Heisenberg
TH Köln (TH)
University of Applied Sciences
Claudiusstr. 1
50678 Köln
Tel.: +49 221 8275 3389
E-Mail:
gernot.heisenberg@th-koeln.de

Dr. Gernot Heisenberg ist Professor für Information Research and Data Analytics an der TH Köln.

Weitere Informationen
www.gernotheisenberg.de

questback

Tina Hees
Questback GmbH
Gustav-Heinemann-Ufer 72a
50968 Köln
Tel.: +49 221 27169 729
E-Mail:
Tina.Hees@questback.com

Tina Hees ist Business Intelligence Developer bei Questback GmbH.

Weitere Informationen
www.questback.de