Optimierung durch Strukturierung

Mit modernen Ansätzen Transparenz schaffen und das Potenzial von unstrukturierten Textdaten nutzen

*Jürgen Fritz**

Obwohl hochauflösende Sensoren die detaillierte Erfassung von numerischen Produkt- und Prozessparametern erlauben, stellen doch Textdaten den größten Anteil aller verfügbaren Daten in den meisten Unternehmen. Genutzt werden diese Daten viel zu wenig. In diesem Beitrag werden aufgezeigt, welche Möglichkeiten linguistische, statistische Techniken und insbesondere Methoden der Künstlichen Intelligenz im Bereich der Textanalyse bieten und wie diese sich konkret anwenden lassen.

Eine große Anzahl an hochauflösenden Sensoren zeichnet ständig numerische Werte zu unterschiedlichsten Parametern auf und lässt so einen großen Datenschatz entstehen. Fotos, Videos und Audioaufnahmen tragen ebenfalls massiv zum Zuwachs der Datenmenge bei.

An Textdaten wird indes häufig nicht direkt gedacht, wenn es um das Thema Data Mining geht. Dabei liegen bis zu 80 Prozent aller Informationen in Textform vor. Zu diesen textuellen Ressourcen zählen Webquellen, Bücher, E-Mails und Artikel, welche zusätzlich immer häufiger in digitaler und damit für Rechner verarbeitbarer Form vorliegen. Von immer größerer Bedeutung sind die sozialen Netzwerke, die viele Anwender zunehmend für Informationszwecke und den Austausch zu unterschiedlichsten Themen verwenden.

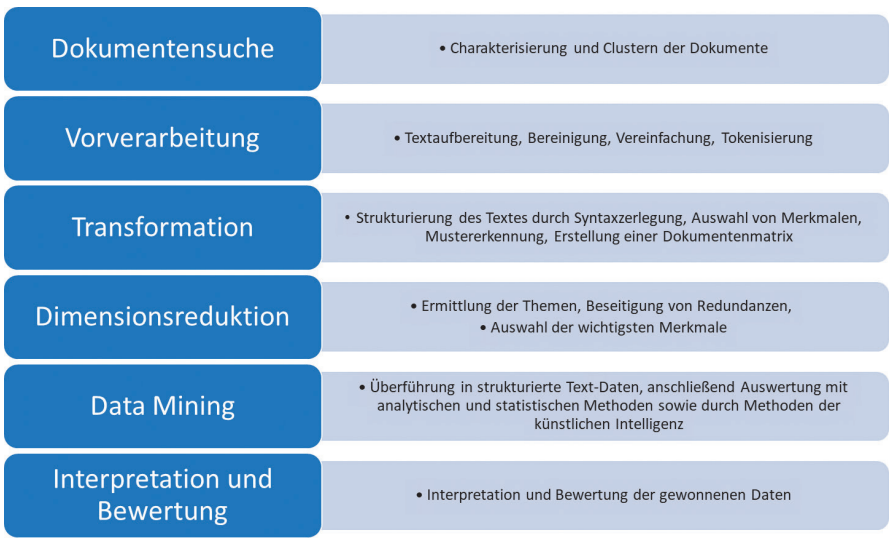

Bild 1. Die wesentlichen Schritte beim Text Mining

Die Textanalyse wird meist auch als Text Mining bezeichnet. Dabei werden linguistische und statistische Techniken sowie Methoden der Künstlichen Intelligenz eingesetzt, um die Informationen aus bislang nicht weiter bekannten Daten zu extrahieren und weiterzuverarbeiten. Text Mining bietet grundsätzlich viele Techniken und Methoden, um den Datenschatz in Textform zu heben und Entscheidungen auf ein datenbasiertes Fundament zu stellen [1].

Beim Text Mining wird ähnlich wie bei der allgemeinen Datenanalyse vorgegangen, allerdings müssen gewisse Besonderheiten mitberücksichtigt werden. Im Einzelnen läuft der Prozess des Text Minings in den folgenden Schritten ab (Bild 1) [2]:

- *Dokumentensuche (Information Retrieval)*
  Charakterisierung und Clusterung der für die weitergehende Analyse relevanten Dokumente;

* **Korrespondenzautor**
Prof. Dr.-Ing. Jürgen Fritz; Head of Industry Liaison and R&D, Member of the Management Board, Haute école de gestion Fribourg (HEG-FR); Chemin du Musée 4, CH-1700 Fribourg;
Tel.: +41 26 429 63 24; E-Mail: juergen.fritz@hefr.ch

Hinweis
Bei diesem Beitrag handelt es sich um einen von den Mitgliedern des ZWF-Advisory-Board wissenschaftlich begutachteten Fachaufsatz (Peer Review).

Bild 2. Beispiel einer Word Cloud (Schlagwortwolke), erstellt mit Python in JupyterLab (Worte werden grafisch dargestellt und hervorgehoben)

- *Vorverarbeitung*
  Zur zielgerichteten Weiterverarbeitung wird der Text aufbereitet, bereinigt und sofern möglich auch vereinfacht. Zu diesem Schritt zählt beispielsweise die Tokenisierung, d. h. die Segmentierung des Textes in Wörter;
- *Transformation*
  Strukturierung des Textes durch Syntaxzerlegung (Parsing), Auswahl von Merkmalen, Mustererkennung, Erstellung einer Dokumentenmatrix;
- *Dimensionsreduktion*
  Themen werden ermittelt, Redundanzen beseitigt und die wichtigsten Merkmale werden ausgewählt;
- *Data Mining*
  Durch die ersten vier Schritte des Text Mining wird der unstrukturierte Text durch Methoden der Verarbeitung natürlicher Sprache (Natural Language Processing – NLP) in strukturierte Textdaten überführt, welche dann mit analytischen und statistischen Methoden sowie zunehmend durch Künstliche Intelligenz weiter ausgewertet werden können;
- *Interpretation und Bewertung*
  Die gewonnenen Daten werden interpretiert und bewertet.

Methoden des Text Mining

Die Analyse mittels Text Mining automatisiert oder unterstützt folgende Aufgaben:

Zusammenfassen
Erstellung einer Zusammenfassung unter Beibehaltung der Kerninformationen von Dokumenten, welche aus vielen Informationen bestehen. Dabei werden die Sätze und Satzteile aus mehreren Dokumenten identifiziert, welche den Inhalt besonders gut repräsentieren.

Informationsextraktion
Daten zu Merkmalen, wie z. B. zu Personen, Adressen, Telefonnummern, E-Mail- Adressen, Orte oder Organisationen, lassen sich aus einem Text entnehmen. Die Anwendungsfälle sind breit und unterschiedlich. Ein Beispiel ist die Strukturierung von Daten aus Kleinanzeigen, die in freiem Text vorliegen. Im Qualitätsbereich können Daten zu Reklamationen, Kundenbewertungen automatisch strukturiert und auswertbar gemacht werden.

Quantitative Analysen
Die Häufigkeit und Verteilung von Satzlängen, Wörtern, Koreferenzen (Verwendung von zwei verschiedenen Ausdrücken für dasselbe Objekt, zum Beispiel: das Auto; es) oder von anderen Mustern können durch quantitative Analysen näher beschrieben werden.

Klassifikation
Zuordnung von Dokumenten in bestimmte Klassen. Die Klassifikation wird beispielsweise in Bibliotheken gemacht. Über die Zuordnung von Merkmalen (Labels) kann die Klassifikation von Dokumenten verfeinert werden.

Clusterisierung
Auf Basis der Ergebnisse der Textanalyse werden Dokumente in Cluster ähnlicher Dokumente gruppiert. Damit kann ein Index erstellt werden, der Suchvorgänge erheblich beschleunigt.

Visualisierung
Informationen zu Dokumenten werden visualisiert, um die Ergebnisse der Analysen klarer darzustellen. Ein Beispiel ist die Schlagwortwolke (englisch: Word Cloud), mit der beispielsweise die Worthäufigkeit hervorgehoben wird (Bild 2).

Themenerkennung
Insbesondere in sozialen Netzwerken ermöglicht der zeitliche Verlauf des Auftretens bestimmter Muster und Begriffe Rückschlüsse auf wichtiger werdende Themen und Trends. Im Rahmen einer aktiven Feldbeobachtung kann so die zeitliche Häufung eines Problems dargestellt werden.

Analyse von thematischer Zusammengehörigkeit und Verknüpfbarkeit
Mögliche Verknüpfungen zwischen den Themen, die in einem Text behandelt werden, können ebenfalls erkannt werden. Mögliche Verknüpfungen sind Wiederholungen, Relationen oder Verkettungen in einem Text. Auf dieser Basis können Markierungen (Tags) und Anmerkungen (Annotations) erstellt werden.

Sentiment- und Meinungsanalysen
Ein Dokument besteht in der Regel nicht nur aus Fakten, sondern auch aus der Meinung des Verfassers (Bild 3). Diese kann positiv, negativ, neutral oder gemischt sein. Deren automatische Erkennung wird als Sentimentanalyse bezeichnet. Das Ziel ist es subjektive Informatio-

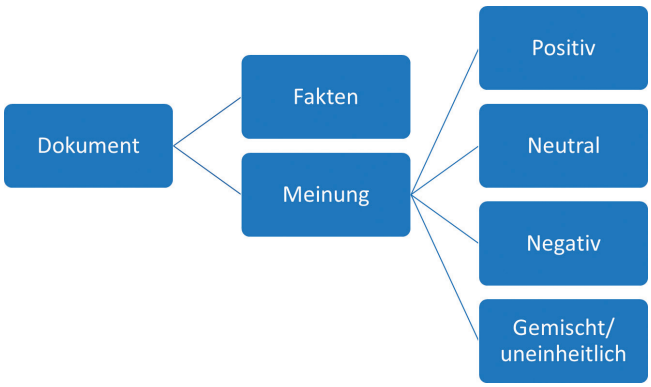

Bild 3. Dokumente bestehen aus Fakten und Meinungen, die positiv, neutral, negativ oder gemischt sein können – Meinungen zu identifizieren ist Ziel von Sentimentanalysen

nen zu objektivieren und Emotionen zu quantifizieren.

In den letzten Jahren wurden in diesem Bereich sowohl in der Forschung als auch bei der Implementierung in Tools erhebliche Fortschritte erzielt. Insbesondere der Einsatz von Methoden der Künstlichen Intelligenz haben die erreichbaren Ergebnisse deutlich verbessert. Sentimentanalysen aus den sozialen Netzwerken können mit Daten aus dem Customer Relationship Management (CRM)-System kombiniert werden. In Systemen wie SAP HANA ist die Möglichkeit Sentimentanalysen der Durchführung von bereits seit einiger Zeit integriert.

Anwendungsmöglichkeiten

Zur systematischen Identifikation von Anwendungsmöglichkeiten von Text Mining in Unternehmen eignet sich die Orientierung an den Geschäftsprozessen [2]. Bild 4 fasst die Anwendungsmöglichkeiten von Text Mining für Management-Prozesse, operative Kernprozesse und Unterstützungsprozesse zusammen.

Für die Entwicklung der Unternehmensstrategie und die Definition von Unternehmenszielen ist es erforderlich, sich einen umfassenden Überblick über die Marktanforderungen einerseits sowie die Kundenforderungen andererseits zu verschaffen. So lassen sich Trends schon frühzeitig identifizieren, wodurch ein wichtiger Vorsprung vor dem Wettbewerb möglich wird. Auch die Analyse des Wettbewerbes selbst wird durch den Einsatz von Text Mining wesentlich umfangreicher und einfacher, da umfassende Datenbestände zielgerichtet analysiert werden können.

Die Kundenforderungen können teilweise durch Umfragen zielgerichtet ermittelt werden. Genauso interessant sind heutzutage jedoch auch die Berichte und Kommentare, welche über soziale Medien ausgetauscht werden. Sentiment-Analysen erlauben es dabei automatisch zu identifizieren, ob die Bewertungen eher positiv oder negativ ausfallen. Dadurch kann etwa zielgerichtet nach Verbesserungspotenzialen gesucht oder eine Wissensdatenbank aufgebaut werden.

Für die Produkt- und Prozessentwicklung ist es entscheidend sicherzustellen, dass möglicherweise bestehende Patente

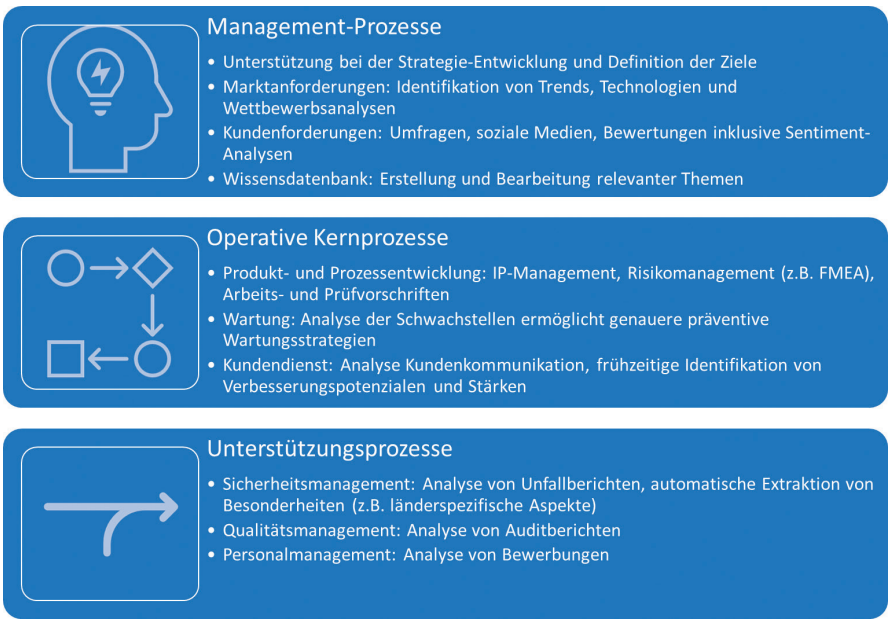

Bild 4. Anwendungsmöglichkeiten von Text Mining in unterschiedlichen Geschäftsprozessen

nicht verletzt und gleichzeitig geeignete Innovationen als Patente angemeldet werden (IP-Management). Im Rahmen des Risikomanagements erlaubt Text Mining die Analyse umfassender Dokumente wie beispielsweise von FMEA (vgl. [3]). Auch Arbeits- und Prüfvorschriften zu bestehenden Produkten können für die Entwicklung neuer Generationen relevante Daten beinhalten.

Durch den Produkteinsatz und in der Prozessanwendung werden wichtige Daten in Form von Wartungsberichten generiert. Auch dabei handelt es sich in der Regel um textuelle Daten. Eine umfassende Analyse erlaubt die systematische Erfassung von Schwachstellen. Dadurch können Produkte und Prozesse verbessert und Wartungsstrategien zielgerichtet optimiert werden.

Auch im Kundendienst fällt kontinuierlich ein wichtiger Datenschatz an. Die Analyse der Kommunikation mit den Kunden erlaubt es Stärken auszubauen und Verbesserungspotenziale zu heben. Mit Text Mining ist es möglich Kundenrückmeldungen zu kategorisieren und automatisiert an die zuständigen Abteilungen weiterzuleiten. Dadurch werden die Bearbeitungsdauer und Kosten reduziert sowie die Kundenzufriedenheit verbessert (vgl. [4]). Wie sich beispielsweise Einsatzberichte von Service-Technikern und andere Kurztexte gewinnbringend analysieren und verwerten lassen wird in [5] gezeigt.

Besonders vielseitig sind die Einsatzmöglichkeiten des Text Mining im Bereich der Unterstützungsprozesse. Durch die Analyse von Unfallberichten kann das Sicherheitsmanagement zielgerichtet und bei globalen Organisationen beispielsweise länderspezifisch verbessert werden.

Im Qualitätsmanagement können Auditberichte analysiert werden. Dadurch können Schwerpunkte von Abweichungen besser festgestellt werden. Mit zielgerichteten Maßnahmen können die Ursachen dann proaktiv beseitigt werden.

In Zeiten des Fachkräftemangels kann das Personalmanagement durch die automatische Voranalyse von Bewerbungen verbessert werden. So können interessante Kandidaten für offene Stellen besser und schneller identifiziert werden. Dies kann entscheidend sein, um die besten Mitarbeiter für das eigene Unternehmen zu gewinnen.

Umsetzung von Text Mining

Für die Umsetzung von Text Mining existieren vielfältige Lösungen. Die Auswahl hängt von der konkreten Aufgabe und deren Zielsetzung ab. Über API (Application Programming Interface) kann die

Text-Mining-Funktionalität in der Regel in bestehende Anwendungen integriert werden.

Die großen Anbieter von Cloud-Diensten bieten mit Microsoft Azure Text Analytics oder Google Cloud Natural Language ausgereifte Lösungen an, die viele der in diesem Artikel vorgestellten Methoden beinhalten. Auch IBM bietet mit Watson eine bekannte und leistungsfähige Plattform an. Darüber hinaus existieren Softwarelösungen wie RapidMiner oder SAS Text Analytics, die ebenfalls in Betracht gezogen werden können.

Auch im Bereich Open Source finden sich interessante Anwendungen. So kann beispielsweise das Natural Language Tool Kit (NLTK) in Python eingebunden und an die jeweiligen Bedarfe angepasst werden.

Angesichts des Hypes der vergangenen Monate kann der Verweis auf ChatGPT in einem Beitrag zu Text Mining natürlich nicht ausbleiben. Das dieser Plattform zugrundeliegende Sprachmodell GPT (Generative Pre-Trained Transformer) kann für die meisten hier vorgestellten Aufgaben angewandt werden – es ist vielmehr zu prüfen, ob es nicht jeweils effizientere Lösungen gibt. Technologisch handelt es sich bei diesem Modell um ein so genanntes Large Language Model (LLM), dessen grundlegende Funktionsweise in [6] gut erklärt wird. Auch Google liefert mit Bard ein grundlegend vergleichbares Modell an.

Neben der technischen Realisierbarkeit müssen in der industriellen Anwendung auch Fragen zur Datensicherheit berücksichtigt werden. Die hier genannten Lösungen basieren indes teilweise auf dem Prinzip der Datenübertragung an Modelle außerhalb des Firmennetzwerkes. Für kritische Daten kann deren Einsatz ein Ausschlusskriterium sein.

Zusammenfassung und Ausblick

Die Möglichkeiten, welche sich durch Methoden des Text Mining bieten, sind vielfältig und umfassend. Viele Lösungen sind direkt nutzbar oder können rasch eingesetzt und integriert werden. Dennoch existieren auch weiterhin Forschungsaktivitäten, um die Durchführungsqualität bestimmter Aufgaben weiter zu erhöhen (vgl. [7]). Außerdem lassen sich Detailkenntnisse zu Fachgebieten oder fortgeschrittene linguistische Kenntnisse nicht durch Text Mining ersetzen.

Insbesondere für fortgeschrittene Methoden des Text Mining unter Anwendung von Künstlicher Intelligenz ist es, wie generell beim Einsatz dieser Technologie, wichtig zu verstehen, wie die Modelle grundsätzlich funktionieren. Die für das Training verwendete Datenbasis beeinflusst die Ergebnisse bei der Modellanwendung entscheidend.

Auch folgende Aspekte müssen den Anwendern nähergebracht werden: Der erfolgreiche und verantwortungsbewusste Einsatz von Text Mining basiert letzten Endes auf den Fähigkeiten der Mitarbeitenden und kann diese unterstützen, jedoch nicht ersetzen.

Literatur

1. Isaenco, A.: Dokumentensammlungen analysieren – Daten schürfen mit Text Mining. IT & Production – Das Industrie 4.0-Magazin für erfolgreiche Produktion (2022), S. 32–34. Online unter https://www.it-production.com/produktionsmanagement/daten-schuerfen-mit-text-mining/ [Abruf am am 16.03.2023] DOI:10.1007/s43443-022-0334-z
2. Fritz, J.: Datenbasierte Optimierung des Business Management Systems – Geschäftsprozesse verbessern mit Data Analytics, Industrie 4.0, KI, Chatbots und Co. Carl Hanser Verlag, München 2022 DOI:10.3139/9783446472549.fm
3. Chen, L.; Nayak, R.: A Case Study of Failure Mode Analysis with Text Mining Methods, In: Ong, K.-L.; Li, W.; Gao, J. (Hrsg.): Proceedings of the 2nd International Workshop on Integrating Artificial Intelligence and Data Mining (AIDM 2007). CRPIT 84 (2007), S. 49–60
4. Kube, B.: Näher am Fluggast: Brussels Airlines optimiert die Auswertung von Kundenfeedback durch KI-Technologien, Lufthansa Industry Solutions. Online unter https://www.lufthansa-industry-solutions.com/ [Abruf am am 16.03.2023]
5. Klingler, N.: Von der Datenablage zur Wissensquelle – Technische Kurztexte in der Automobilentwicklung. Online unter https://www.elektroniknet.de/automotive/software-tools/technische-kurztexte-in-der-automobilentwicklung.200117.html [Abruf am 15.03.2023]
6. Wolfram, S.: What is ChatGPT Doing... and Why Does it Work? Online unter https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/ [Abruf am 19.03.2023]
7. Yücel, A.; Dag, A.; Oztekin. A.; Carpenter, M.: A Novel Text Analytic Methodology for Classification of Product and Service Reviews. Journal of Business Research 151 (2022), S. 287–297 DOI:10.1016/j.jbusres.2022.06.062

Der Autor dieses Beitrags

Prof. Dr.-Ing. Jürgen Fritz, geb. 1979, ist Mitglied des Direktoriums der Hochschule für Wirtschaft in Freiburg, Schweiz, welche zur HES-SO Fachhochschule Westschweiz zählt. Er ist Leiter für industrielle Kontakte sowie Forschung und Entwicklung und hat über 15 Jahre internationale Führungsverantwortung in der Forschung und Vorausentwicklung, der Produktentwicklung, der Produktion, in der Qualität und im Einkauf.

Abstract

Optimization through Structuring – Modern Approaches to Create Transparency and Leverage the Potential of Unstructured Text Data. Although advanced sensors allow the detailed recording of numerical product and process parameters, textual data represents the largest share of all available data in most companies. Nevertheless, this data is used far too little. What possibilities linguistic, statistical and especially methods of artificial intelligence offer in the field of text analysis and how they can be applied in practice will be demonstrated in this article

Schlüsselwörter

Text Mining, Künstliche Intelligenz, Qualitätsmanagement, Business Management System, Industrielle Anwendungen, Large Language Models (LLM)

Keywords

Text Mining, Artificial Intelligence, Quality Management, Business Management System, Industrial Applications, Large Language Models (LLM)

Bibliography

DOI:10.1515/zwf-2023-1078
ZWF 118 (2023) 6; page 432 – 435

ISSN 0947-0085 · e-ISSN 2511-0896