technische universität
dortmund

Fachwissenschaftliche Projektarbeit (MA)

Analyse und Aufbereitung des Themenfeldes
Text Mining für Studierende

André Stade
Matrikelnummer: 150359
Studiengang: Logistik M.Sc.

ausgegeben am:
15.01.2019

eingereicht am:
30.07.2019

Betreuer:
Prof. Dr.-Ing. Markus Rabe
Dr.-Ing. Anne Antonia Scheidler

Technische Universität Dortmund
Fakultät Maschinenbau
Fachgebiet IT in Produktion und Logistik
www.itpl.mb.tu-dortmund.de

Inhaltsverzeichnis

1 Einleitung

Die Gewinnung von Wissen aus Textdokumenten hat in den letzten Jahren stark an Bedeutung gewonnen. Zum einen ist der Grund dafür die immer schneller waschende Datenmenge. Der Hersteller von Speichermedien Seagate prognostiziert, dass das weltweite Speicheraufkommen sich bis zum Jahr 2025 auf 175 Zettabytes mehr als verfünffacht; für das Jahr 2018 wurde es auf 33 Zettabytes geschätzt [Reinsel et al. 2018]. Weiterhin wird angenommen, dass mehr als 80% des geschäftsrelevanten Wissens in Form von unstrukturierten Informationen gespeichert ist [Pfeifer 2014, S.1]. Wissen ist der wesentlichste Faktor wirtschaftlicher Wertschöpfungsketten. Es ist für alle Bereiche relevant von der Forschung und Entwicklung bis hin zum Vertrieb [Heyer et al. 2012, S.1].

Ein weiterer Grund für die wachsende Bedeutung der Verarbeitung von Text ist die zunehmende Interaktion im Internet durch E-Commerce und Social Web [Walsh und Möhring 2014, S.70]. Die meisten Daten werden in Form von Textdaten generiert wie z. B. auf Internetseiten, in Office-Dokumenten oder in Foreneinträgen. Um diese unstrukturierte Datenmenge nutzen zu können, bedarf es spezialisierter Computeranwendungen [Pfeifer 2014, S.1].

Bevor unstrukturierte Informationen verarbeitet werden können, müssen sie im Gegensatz zu einer Datenbank, vorher strukturiert werden. Dieser Vorgang ist mit hohem Aufwand verbunden [Heyer et al. 2012, S.3].

Diese Arbeit soll *Text Mining* als computergestütztes Verfahren näher erläutern, mit dessen Hilfe semantische Analysen von Texten durchgeführt werden können. Dabei wird zuerst eine begriffliche Einordnung des Terminus vorgenommen. Anschließend werden Verfahren aufgezeigt, die notwendig sind, um Dokumente verarbeiten zu können.

Danach wird auf Besonderheiten von Sprache im Hinblick auf statistische Eigenschaften eingegangen und wie sich diese für das Text Mining nutzen lassen können. Anschließend erfolgt eine Vorstellung der gebräuchlichsten Methoden, um je nach Informationstyp das Text Mining zu unterstützen. Anknüpfend wird IBM Watson als Computersystem für die Verarbeitung von natürlicher Sprache vorgestellt.

Abschließend folgen Praxisbeispiele aus dem Themengebiet der Produktion und Logistik. Dabei werden Beispiele aus dem Bereich des E-Commerce sowie der Produktentstehung vorgestellt.

2 Text Mining

Dieses Kapitel beschäftigt sich zunächst mit den verschiedenen Definitionsansätzen des Text Minings und den unterschiedlichen Sichtweisen. Anschließend wird ein allgemeines Vorgehen aufgezeigt, durch welches sich Textdokumente automatisiert verarbeiten lassen.

Auffälligkeiten und Besonderheiten der Sprache folgen im Anschluss und wie sich diese für die Analyse nutzen lassen. Beispiele für Verfahren des Text Minings werden danach kurz vorgestellt sowie eine Umsetzung in Form von IBM Watson.

2.1 Definition des Begriffs „Text Mining“

Unter Text Mining wird die Gewinnung von Wissen aus schwach- oder unstrukturierten Textdaten verstanden. In Tabelle 2.1 wird eine Abgrenzung von anderen verwandten Themengebieten vorgenommen. Dabei wird differenziert nach der Strukturiertheit der Daten in strukturierte (z. B. Datenbanken) und unstrukturierten Daten (z. B. Textdokumente). Zusätzlich wird unterschieden, auf welche Weise das gewünschte Wissen erhalten wird: durch reinen Zugriff (Information durch Abfrage) oder aber durch Interpretation (Wissen durch Anfrage) [Pfeifer 2014, S.17f].

Information Retrieval sowie Text Mining arbeiten beide mit unstrukturierten Daten. Information Retrieval ist dabei das reine Finden von Informationen bzw. Dokumenten, welche Fragestellungen beantworten. Dabei ist explizit das reine Auffinden bzw. Abrufen gemeint. Das Problem ist die Menge an Daten, welche die benötigten Informationen enthält sowie eine Vielzahl an aktuell nicht relevanten Daten. Diese könnten jedoch zu einem späteren Zeitpunkt noch von Interesse sein.

Text Mining ist im Gegensatz dazu explizit die Entdeckung von Wissen. Dabei werden die unstrukturierten Quelldaten interpretiert und Wissen auf diese Art erschlossen [Pfeifer 2014, S.17f; Heyer et al. 2012, S.3f].

In der Literatur wird das Themengebiet des Text Minings zum ersten Mal im Jahr 1995 unter dem Namen „knowledge discovery from text“ erwähnt [Feldman und Dagan 1995]. Dabei wurde bereits auf die Abgrenzung zum Data Mining eingegangen und der Beson-

Tab. 2.1: Abgrenzung wissensgewinnender Disziplinen [Heyer et al. 2012]

	Zugriff Information durch Abfrage	**Interpretation** Wissen durch Anfrage
Strukturierte Daten	Datenbanksysteme	Data Mining
Unstrukturierte Daten	Information Retrieval	Text Mining

derheit, dass sich Text Mining auf unstrukturierten Daten bezieht. Diese werden während des Prozesses Text Mining in eine strukturierte Form überführt [Pfeifer 2014, S.18].

Es existiert keine generell akzeptiere Definition des Begriffs „Text Mining“. Es existieren mehrere Definitionen, die sich zum Teil stark in ihrem Inhalt unterscheiden [Heyer et al. 2012, S.4; Pfeifer 2014, S.18]. Im Folgenden wird ein kurzer Überblick über verschiedene Definitionen gegeben, auf die [Hotho et al. 2005] referenzieren:

- Text Mining = Informationsextraktion [Sebastiani 2002]

  Unter Text Mining versteht Sebastiani hauptsächlich den Prozess zur Extraktion von Informationen aus Texten: „*Text Mining is increasingly being used to denote all the tasks that, by analyzing large quantities of text and detecting usage patterns, try to extract probably useful (although only probably correct) information.*“ [Sebastiani 2002, S.2]

- Text Mining = Text Data Mining [Kosala und Blockeel 2000]

  Der Definitionsansatz von Kosala und Blockeel setzt den Fokus auf die Erkennung von Mustern in Texten mit der Hilfe von Methoden und Algorithmen. Diese können dabei nicht nur aus den Bereichen der Informationsextraktion und natürlichen Sprachverarbeitung stammen, sondern auch aus Teilbereichen des Machine Learning.

- Text Mining = Prozess der Wissensgewinnung aus Daten [Hearst 1999]

  Hearst versteht unter Text Mining vor allem die Gewinnung von neuem Wissen: „*Another way to view text data mining is as a process of exploratory data analysis that leads to the discovery of heretofore unknown information, or to answers for questions for which the answer is not currently known.*“ [Hearst 1999, S.5] Im Folgenden schließt er dabei Informationsextraktion und Textkategorisierung ausdrücklich aus und ordnet dies dem reinen Informationszugriff zu.

Mehler und Wolff gliedern die verschiedenen Sichtweisen auf das Text Mining schlussendlich in die zwei Grundpostionen:

- Methodische Perspektive
- Wissensorientierte Perspektive

Dabei werden vor allem die zwei Grundpositionen Methodenorientierung mit Wissensorientierung gegenübergestellt. Als unteres Ende werden beim Text Mining Begriff methodenorientierte Ansätze verstanden. Diese analysieren mit welchen Methoden der Textanalyse gewisse Aufgaben mit Erfolg gelöst werden können. Dabei geht es um die Ergänzung, Erweiterung oder Ersetzung der Informationsextraktion [Mehler und Wolff 2005, S.5].

Im Gegensatz dazu existieren am oberen Ansätze wie von Hearst, bei dem ein System selbständig neues Wissen gewinnt. Hier ist von einer wissensorientierten Sichtweise die Rede. Mehler und Wolff kritisieren allerdings, dass bei diesem Ansatz eher von explorativer Textdatenanalyse als von Text Mining gesprochen werden müsse, da kaum erprobte Verfahren existieren [Mehler und Wolff 2005, S.5f].

In Anlehnung an den methodenorientierten Ansatz von Mehler und Wolff wird im Folgenden die Definition von Pfeifer verwendet:

„Text Mining ist die Gesamtheit aller Methoden und Algorithmen zum (halb-) automatisierten Gewinnen von Wissen aus Textdaten“ [Pfeifer 2014, S.19].

Dabei werden verschiedene Textextraktions- und Analysemethoden berücksichtigt und die Informationen müssen nicht unbedingt neuartig sein.

2.2 Allgemeines Vorgehen zur Verarbeitung von Textdokumenten

Anknüpfend an die Definitionsansätze in Abschnitt 2.1 wird in diesem Abschnitt das typische Vorgehen beschrieben, um Dokumente mittels Text Mining verarbeiten zu können. Im Folgenden wird unter Text die natürliche Sprache verstanden und nicht etwa Quelltext oder mathematische Gleichungen, obwohl auch diese Gebiete vom Text Mining abgedeckt werden können.

Zunächst wird beschrieben welche Art von Texten mittels Text Mining verarbeitet werden können. Nachfolgend werden Methoden aufgezeigt, um die Datenquellen zu verarbeiten und schlussendlich Wissen aus Text generieren zu können.

Text Komponenten

Text ist definiert als unstrukturierte Datenmenge aus einzelnen Zeichenketten (engl.: Strings), die Wörter genannt werden. Eine Vielzahl einzelner Zeichenketten (Wörter) muss im Kontext mit den übrigen aus dem Text verstanden werden. Es sind weiterhin Regeln notwendig, um aus mehreren Wörtern einen Text zu generieren; diese wird Grammatik genannt [Jo 2019, S.3]. Die Bezeichnung „unstrukturiert“ bezüglich des Datenformats kann unter Umständen irreführend sein, da die linguistische Struktur sehr wohl durch Grammatik vorgegeben ist [Feldman und Sanger 2008, S.3].

Ein Text ist das Gefüge von einzelnen Einheiten, die durch Grammatik zu einem Satz kombiniert werden. Mehrere Sätze werden zu einem Paragrafen zusammengefasst. Die einzelnen Gefüge sind also: Wörter, Sätze und Paragrafen. Ein Wort wird dabei als kleinste Einheit verstanden, da es im Gegensatz zu einzelnen Buchstaben bereits Bedeutung besitzt, wenn auch losgelöst vom Kontext. Des Weiteren existieren Wörter, die nur eine grammatische Funktion haben und keinen Inhalt wie z. B.: ein, der, die, das. Aus diesem

Grund werden diese Stoppwörter (engl.: Stop-Words) im weiteren Verlauf herausgefiltert [Jo 2019, S.3].

Der Grammatik folgend beginnt jeder Satz mit einem Großbuchstaben und endet mit fest definierten Zeichen. Einzelne Wörter werden mit einem Leerzeichen unterteilt, was Tokenisierung (engl.: Tokenization) genannt wird [Jo 2019, S.5]. Alle relevanten Regeln müssen folglich auch zur Verarbeitung von Texten beachtet werden, damit nur relevante Informationen im Prozess übrig bleiben.

Datenformate

Es existieren verschiedene Datenformate, in denen Text vorliegen kann. Zum einen können dies Dateitypen sein, in denen bestimmte Programme speichern (z. B.: Microsoft Word mit .docx) oder aber standardisierte Formate wie das Portable Document Format (PDF). Des Weiteren existiert einfacher Text (engl.: Plain Text) als simpelstes Format (.txt), da dort keine Formatierung möglich ist. Zu den unstrukturierten Dokumenten gehören Textdokumente, Memos, E-Mails, RSS-Feeds, Blogeinträge, Kurznachrichten (z. B.: Twitter), Forenbeiträge sowie Kommentare in sozialen Netzen.

Zusätzlich dazu existieren semi-strukturierte Datenformate wie Extensible Markup Language (XML) oder Hypertext Markup Language (HTML). XML ist ein flexibles Datenformat, welches auf HTML basiert. Einzelne Felder werden mit Start- sowie Endtags definiert und die eingeschlossenen Texte können somit inhaltlich einem Bereich zugeordnet werden [Jo 2019, S.6; Feldman und Sanger 2008, S.3].

Eine Sammlung unterschiedlicher Texte, die zu einem Themengebiet gehören wird Korpus bzw. Textkorpus genannt. Dies kann z. B. eine Sammlung von Zeitungsartikeln darstellen [Jo 2019, S.6]. Große deutschsprachige Textkorpora-Datenbanken werden z. B. von der Universität Leipzig[1] oder der Berlin-Brandenburgischen Akademie der Wissenschaften[2] angeboten.

In den Datenbanken liegen die Texte häufig bereits als XML vor, sodass zusätzliche XML-Tags nützliche Informationen liefern können wie z. B.: <DATE>, <TITLE>, <SUBJECT>, <TOPIC>, <HEADLINE> und <BODY> [Weiss et al. 2015, S.15]. In Abbildung 2.1 ist beispielhaft ein Text in XML-Darstellung abgebildet.

Text Indexing

Unter Text Indexing wird der Vorgang verstanden, bei dem Text oder Texte in eine Liste von Wörtern überführt werden. Dafür sind mehrere Schritte notwendig, da eine unstrukturierte Datenmenge nicht direkt transformiert werden kann. Für das Indexing werden folgende Schritte durchgeführt:

- Tokenisierung (Tokenization)
- Wortstammreduktion (Stemming)
- Stoppwörter Entfernung (Stop-Word Removal)

[1] `https://wortschatz.uni-leipzig.de`
[2] `https://www.dwds.de`

```
<DOC>
  <TEXT>
    <TITLE>
    Solving Regression Problems with Rule-based Classifiers
    </TITLE>
    <AUTHORS>
      <AUTHOR>
      Nitin Indurkhya
      </AUTHOR>
      <AUTHOR>
      Sholom M. Weiss
      </AUTHOR>
    </AUTHORS>
    <ABSTRACT>
    We describe a lightweight learning method that induces an ensemble of decision-rule
        solutions for regression problems. Instead of direct prediction of a continuous
        output variable, the method discretizes the variable by k-means clustering and
        solves the resultant classification problem. Predictions on new examples are made
        by averaging the mean values of classes with votes that are close in number to the
         most likely class. We provide experimental evidence that this indirect approach
        can often yield strong results for many applications, generally outperforming
        direct approaches such as regression trees and rivaling bagged regression trees.
    </ABSTRACT>
  </TEXT>
</DOC>
```

Abb. 2.1: Beispieltext als XML dargestellt [Weiss et al. 2015, S.16]

Die Tokenization dient dazu den Text durch Leerzeichen oder Interpunktion in einzelne Wörter zu unterteilen. Anschließend wird mit dem Stemming durch grammatikalische Regeln jedes Wort auf seine Ursprungsform zurückgeführt. Der letzte Schritt ist das Entfernen überflüssiger Wörter (sog. Stoppwörter) wie Artikel, Konjunktionen und Präpositionen, da diese keine inhaltliche Bedeutung besitzen und nur für die Grammatik notwendig sind.

Die Tokensierung muss zuerst durchgeführt werden, beim Stemming und Stop-Word Removal kann die Reihenfolge verändert werden. Anschließend können noch weiterführende Aufgaben erfolgen wie das Part-of-Speech-Tagging oder Term Weighting erfolgen [Jo 2019, S.19].

Tokenization Die Tokensierung ist definiert als die Segmentierung von Text oder Texten in einzelne Wörter. Dabei werden Trennzeichen der jeweiligen Sprache wie Leerzeichen oder Interpunktion genutzt. Die einzelnen Token werden als Ergebnis in eine Liste überführt. Zusätzlich werden Sonderzeichen und Zahlen entfernt sowie die einzelnen Wörter in Kleinbuchstaben umgewandelt. Token, die ein oder mehr Sonderzeichen enthalten (z. B.: 16%) werden vollständig entfernt. Am Ende des Vorgangs werden alle doppelten Einträge gelöscht, um Redundanz zu verhindern [Jo 2019, S.21]. Je nach Art des Sonderzeichens können zusätzliche Regeln definiert werden, ob es in der vorliegenden Sprache als Trennzeichen behandelt werden soll oder nicht [Weiss et al. 2015, S.17].

Anknüpfend an die Erstellung einer Liste einzelner Token kann ein komplexeres Konstrukt erzeugen werden die sogenannten N-Gramme. Dabei werden aufeinander folgende

Token miteinander verbunden, um eine Reihenfolge abbilden zu können. Diese Technik wird unter anderem dazu benutzt, um die Wahrscheinlich zu bestimmen, ob ein korrekter Satz vorliegt oder nicht [Heyer et al. 2012, S.102–104].

Stemming Beim Stemming wird jeder zuvor segmentierte Token auf seine grammatikalische Ursprungsform gebracht. Dafür werden Regeln auf die Liste der zuvor ermittelten Tokens angewendet. Üblicherweise werden Substantive, Verben und Adjektive verarbeitet. Die Zielausgabe ist eine bereinigte Liste mit den Stammformen aller Token, bei der die Duplikate entfernt wurden [Jo 2019, S.23].

Substantive im Plural werden auf ihre Form im Singular angepasst. Häufig können dabei im englischen Wortanhängsel wie „s“ oder „es“ weggelassen werden, im Deutschen ist dieser Vorgang wesentlich schwieriger aufgrund der Vielzahl an Sonderregeln. Bevor dieser Vorgang auf Substantive angewendet werden kann, müssen mittels Part-of-Speech-Tagging die einzelnen Wortarten bestimmt werden. Die Verben werden in die jeweilige Form des Infinitivs zurückgeführt. Dabei sind Bestandteile der Konjugation und zeitspezifische Änderungen zu entfernen.

Unterschieden wird zusätzlich zwischen „hartem“ und „weichem“ Stemming. Entscheidend ist hierbei der Grad der Rückführung des Wortes. So kann z. B. „Kategorisierung“ zurückgeführt werden auf „Kategorie“ oder „kategorisieren“ [Weiss et al. 2015, S.19–21; Jo 2019, S.23f].

Stop-Word Removal Während des Stop-Word Removal werden Wörter aus den Listen der Tokenization bzw. des Stemmings entfernt. Dabei handelt es sich um grammatikalische Wörter, die für den Kontext des Texts irrelevant sind. Bei diesem Vorgang werden die Wörter aus der erstellten Liste der Tokens mit einer Liste verglichen, die Stoppwörter der jeweiligen Sprache enthält. Die Übereinstimmungen werden dann aus der Liste der Tokens entfernt. Bei Stoppwörtern handelt es sich vor allem um Präpositionen („in“, „auf“, „zu“, ...), Konjunktionen („und“, „oder“, „aber“, ...) und Artikeln („der“, „die“, „das“, ...). Diese Wörter sind die am häufigsten gebrauchten in jeglicher Art von Textsammlung. Daher führt das Entfernen zu einer erheblichen Verbesserung der Effizienz.

Alternativ zur Nutzung einer Liste von Stoppwörtern, kann die Entfernung noch mit einem anderen Ansatz erreicht werden. Dabei wird ein Verfahren verwendet, das im Abschnitt Term Weighting noch genauer erläutert wird. Es werden dabei die Häufigkeiten der Vorkommnisse von Wörtern in Texten analysiert. Besonders häufig verwendete Wörter kommen sowohl in den zu untersuchenden Texten vor, als auch in anderen. Diese Wörter sind ebenfalls zu entfernen, da sie nicht brauchbar sind, um den Inhalt des Texts zu erfassen [Jo 2019, S.24f].

Part-of-Speech Tagging Sobald der zu analysierende Text in Form von Token vorliegt, können weitere Verfahren angewendet werden. Falls anschließend linguistische Untersuchungen erfolgen sollen, ist das Part-of-Speech (POS) Tagging ein geeignetes Verfahren. Mit dessen Hilfe werden die grammatikalischen Formen der Token bestimmt. In jeder

Sprache gibt es unterschiedliche Besonderheiten der Grammatik, daher muss für die Sprache des Texts ein geeignetes Tagset vorliegen. In diesem wird definiert, welche Wortarten differenziert werden sollen [Weiss et al. 2015, S.30f]. In Deutschland ist das am häufigsten verwendete Set das Stuttgart-Tübingen-Tag-Set (STTS) [Heyer et al. 2012, S.54].

In fast jeder Sprache existieren Bestandteile wie Substantive und Verben; die Anzahl aller Kategorien variiert jedoch stark je nach Sprache. Auch die Anzahl der pro Sprache zu analysierenden grammatikalischen Wortformen kann je nach Zweck angepasst werden und so zusätzlich die Granularität der Analyse bestimmt werden [Weiss et al. 2015, S.30f].

Das Ergebnis des POS Tagging ist annotierter Text, bei dem die einzelnen Wortformen der vorgegebenen Kategorien klassifiziert wurden. Häufig kommen dabei POS-Tags zum Einsatz z. B. in Form von XML-Tags wie in Abbildung 2.2 dargestellt [Heyer et al. 2012, S.53].

Die Schwierigkeit besteht drin, dass gewisse Wörter in ihrer vorliegenden Form in mehrere Kategorien eingeordnet werden können. Daher muss das POS-Tagging die Verwendung in jedem Satz einzeln prüfen.

- Beispielsatz: „Die Sonne scheint in Leipzig.“
  - *Die*: bestimmter Artikel (ART)
  - *Sonne*: Nomen (NN)
  - *scheint*: Verb (VVFIN)
  - *in*: Präposition (APPR)
  - *Leipzig*: Eigennamen (NE)

```
<Satz>
  <ART>Die</ART> <NN>Sonne</NN> <VVFIN>scheint</VVFIN> <APPR>in</APPR> <NE>Leipzig</NE>
</Satz>
```

Abb. 2.2: Beispieltext POS-Tagging in XML-Darstellung [Heyer et al. 2012, S.53]

Term Weighting Mit Hilfe des Term Weighting werden Wörter gewichtet und damit deren Relevanz für den Text bestimmt. Es kommen dabei die Verfahren Term Frequency (TF) und insbesondere Term Frequency-Inverse Term Frequency (TF-IDF) zum Einsatz. Zuvor sollten bereits die Stoppwörter entfernt worden sein, da diese Gruppe an Wörtern am häufigsten in Texten enthalten ist.

Beim einfachen TF werden die Häufigkeiten der Wörter des vorliegenden Texts bestimmt, indem die Anzahl der Vorkommnisse jedes einzelnen Worts gezählt werden. Es wird differenziert zwischen absoluter und relativer Häufigkeit, welche die Gesamtanzahl der Wörter in Bezug setzt. Die relative Häufigkeit wird bevorzugt eingesetzt, da die Textlänge dort Berücksichtigung findet und keine Verfälschung erfolgt. In der Praxis wird das TF jedoch nur als Vorbereitung für das folgende Verfahren benutzt.

Bei dem TF-IDF werden dagegen alle Texte eines Themengebietes (Textkorpus) berücksichtigt und die Häufigkeiten der Vorkommnisse aller Wörter dieser Texte mit dem

Tab. 2.2: Sortierung nach Häufigkeit [Heyer et al. 2012, S.88]

Rang r	Wortform	Häufigkeit n (Mio.)	r^*n (Mio.)
1	der	7,378	7,378
2	die	7,036	14,072
3	und	4,813	14,440
4	in	7,769	15,074
5	den	2,717	13,586
6	von	2,251	13,504

aktuell Vorliegenden verglichen. Die Gewichtung der einzelnen Wörter nach dem TF-IDF-Schema ist proportional zu den Ereignissen im gegebenen Text, aber umgekehrt proportional zu den anderen Texten des Textkorpus [Jo 2019, S.25f].

Falls ein Wort in vielen Dokumenten erwähnt wird, gilt es als weniger wichtig und die Skala des TF-IDF wird gesenkt; eventuell sogar gegen Null. Wird das Wort jedoch wenig Texten benutzt und es ist eindeutig, erfolgt eine Anpassung des Skalierungsfaktors nach oben, da das Wort als „wichtig“ bewertet wird [Weiss et al. 2015, S.25].

2.3 Besonderheiten bei der Anwendung von Text Mining

Texte stellen keine zufällige Anordnung von Wortformen dar, sondern diese werden durch Regeln der Sprache verbunden, was in Form von syntaktischen, stilistischen und rhetorischen Kombinationen erfolgt. Durch musterbasierte Analysen lassen sich bestimmte Grundmuster innerhalb der Sprache erkennen; in diesem Zusammenhang wird von Sprachstatistik gesprochen [Heyer et al. 2012, S.85].

Zipfsche Gesetze

Mit Hilfe des Zipfschen Gesetzes sind Aussagen über die Häufigkeit und den Umfang des Vokabulars eines Textes möglich. Auch die Art der Veränderung bei Erhöhung der Textmenge lässt sich abschätzen. Auch eine Ermittlung der Mindestgröße der Textemenge für anschließende Analysen sind möglich.

Nach George K. Zipf folgt natürliche Sprache dem *Prinzip der geringsten Anstrengung*. Aus diesem Grund sind die am häufigsten benutzten Wörter kurze Funktionswörter. Werden Wortformen wie in Tabelle 2.2 nach der absoluten Häufigkeit absteigend sortiert, ist Gesetzmäßigkeit von Zipf abzulesen [Heyer et al. 2012, S.87f].

Rang r einer Wortform in der Liste multipliziert mit dessen Häufigkeit n ist in etwa konstant (für $r \geq 2$): $r \cdot n \approx k$. Andersherum ist die Häufigkeit des Vorkommens einer Wortform näherungsweise negativ proportional zu dessen Rang: $n \sim \frac{1}{r}$ [Heyer et al. 2012, S.88].

Diese Verhältnismäßigkeit wurde mehrfach bestätigt für große Textsammlungen wie z. B. Reuters-RCV1[3] und Projekt Wortschatz-Lexikon[4] [Christopher Manning et al. 2009, S.90; Heyer et al. 2012, S.89].

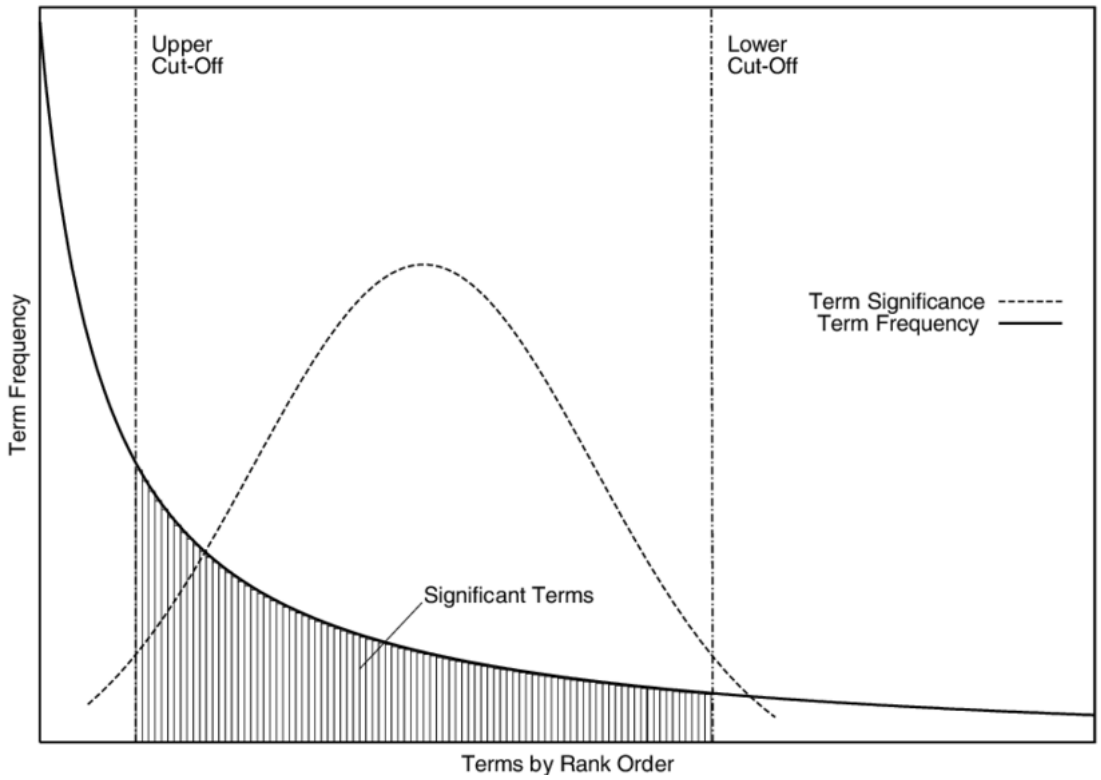

Abb. 2.3: Luhns Anwendung des Zipfschen Gesetzes [Lanquillon 2001, S.37]

Luhns Gesetz

Luhn hat die Überlegungen von Zipf weitergeführt, sodass sich dessen beschriebene Gesetzmäßigkeit für die Anwendung des Text Minings nutzen lässt. Dabei wird die Häufigkeit des Auftretens von Wörtern in Texten dazu genutzt, um die Bedeutung der Wörter für den Inhalt des Texts zu messen. Weiterhin ist die relative Position eines Wortes innerhalb eines Satzes dafür ausschlaggebend, wie relevant dieses für den Kontext ist.

Des Weiteren beschreibt Luhn, dass für die automatische Textanalyse die Auftrittshäufigkeit von Wörtern genutzt werden kann, um einzelne Wörter und Sätze aus Texten zu extrahieren, die das jeweilige Dokument repräsentieren [Rijsbergen 1979, S.10].

In Abbildung 2.3 ist Luhns Ansatz dargestellt. Der gestrichelte Graph *(Term Frequency)* stellt dabei das Zipfsche Gesetz dar. Druchgezogen gezeichnet *(Term Significance)* ist die von Luhn formulierte Auffälligkeit in bezüglich der Relevanz der Wörter.

Es ist möglich, dass die obersten sowie untersten Worthäufigkeiten der Verteilung entfernt werden können, ohne relevante Inhalte zu verlieren. Wörter im oberen Schnittbereich werden als sehr gewöhnlich und nicht signifikant betrachtet. Im unteren Bereich dagegen

[3] `https://trec.nist.gov/data/reuters/reuters.html`
[4] `https://wortschatz.uni-leipzig.de`

werden Wörter verwendet, die so selten benutzt werden, dass sie ebenfalls keine Relevanz für den Text besitzen.

Zwischen den beiden Schnittpunkten verbleiben die relevanten Wörter, dargestellt durch *Term Significance* sowie gestrichelt unterhalb des Graphens *Term Frequency*. Dabei ist die Relevanz links und rechts an den Schnittpunkten nahe Null und in der Mitte am höchsten. Die Schwierigkeit besteht darin, die Schnittpunkte korrekt zu ermitteln, was durch „Trial and Error“ erfolgt [Rijsbergen 1979, S.11].

2.4 Beispiele für die Anwendung von Text Mining

Text Mining Methoden

Für das Text Minig wurden eine Vielzahl von verschiedenen Methoden entwickelt. Diese unterstützen den Prozess des Text Minings je nach Informationstyp. Aufgrund der fehlenden einheitlichen Definition von Text Mining und der stetigen Weiterentwicklung wird im Folgenden lediglich eine kurze Aufzählung über die gebräuchlichsten Methoden gegeben [Pfeifer 2014, S.19–21]:

- Informationsextraktion (Information Extraction)

  Die gezielte Suche nach Informationen in Textdokumenten und deren Überführung in eine strukturierte Form wird Informationsextraktion genannt. Dabei werden konkrete Textstellen gefunden und diesen definierten Bereichen zugeordnet wie Personen oder Orte.

- Kategorisieren (Categorize)

  Ein Textdokument wird analysiert und basierend auf dessen Inhalt werden eine oder mehrere Kategorien zugewiesen. Es kann so die Abgrenzung verschiedener Themengebiete in Nachrichtentexten erfolgen (Sport, Finanzen, Politik, ...).

- Clusterbildung (Clustering)

  Die Bildung von Clustern erfolgt im Gegensatz zum Kategorisieren automatisch. Dabei werden die Themen der Cluster nicht vorher definiert, sondern werden während des Prozesses generiert.

- Stichwortextraktion (Keyword Extraction)

  Für den Text wesentliche Stichwörter werden erkannt und extrahiert. Diese sollen den Inhalt des Texts möglichst gut widerspiegeln.

- Identifikation von Konzepten (Concept Tagging)

  Aus dem Inhalt der Textdokumente werden Konzepte abgeleitet und die Texte werden diesen anschließend zugeordnet. Der Unterschied zur Stichwortextraktion ist, dass diese Konzepte nicht unmittelbar im Text enthalten sein müssen.

- Verbindung von Themen (Concept Linkage)

  Hierbei werden zwischen verwandten Textdokumenten Zusammenhänge hergestellt, die auf den Themen basieren. Herkömmliche Suchmethoden liefern bei diesem Prozess häufig kein gutes Ergebnis und Verbindungen werden nur selten oder gar nicht gefunden. Ein Anwendungsfall wäre die Verbindung von Krankheiten zu möglichen Behandlungsmaßnahmen, die wegen der Datenmenge nicht manuell identifizierbar sind.

- Themen-Verfolgung (Topic Tracking)

  Es werden Schlüsselwörter definiert, mit denen Informationen aus dem Internet abgefragt werden. Falls zu einem Schlüsselwort neue Daten bzw. Themen verfügbar sind, erfolgt eine automatische Benachrichtigung.

- Sentimentanalyse (Sentiment Analysis/Opinion Mining)

  Texte können gezielt Meinungen oder Stimmungen vermitteln. Mittels der Sentimentanalyse können subjektive Informationen aus Textdokumenten ermittelt und die Polarität von Texten bestimmt werden.

- Zusammenfassen (Summarization)

  Die automatisierte Zusammenfassung von einem oder mehreren Texten. Dabei soll der Inhalt möglichst exakt erkannt und anschließend wiedergeben werden.

- Informationsvisualisierung (Information Visualizing)

  Eine Textmenge wird visuell dargestellt; in Form von Karten oder hierarchischen Abbildungen. Bei der Betrachtung einzelner Themengebiete kann die Skalierung der Teilbereiche angepasst werden.

IBM Watson

Die Firma IBM hat 2007 damit begonnen ein Computersystem zu entwickeln, welches offene Fragestellungen gut genug verarbeiten kann, um mit den besten Spielern der Quizshow *Jeopardy!* konkurrieren zu können. Im Jahr 2011 schlug das entwickelte System „Watson“ mithilfe von DeepQA die beiden Spieler mit dem höchsten Ranking in einer Show. Dafür wurde 4 Jahre lang das Watson System entwickelt mit der darunterliegenden DeepQA Antworttechnologie [Gliozzo et al. 2013, S.85].

Die Schwierigkeit für Watson ist dabei, dass ein kurzer Text als Antwort gefordert ist und nicht etwa eine Liste mit relevanten Dokumenten. Weiterhin muss die Fragestellung exakt verstanden werden, welche häufig mehrere Bestandteile enthält. Ebenso muss eine präzise Wahrscheinlichkeit ermittelt werden, mit der Watson davon ausgeht, dass es die korrekte Antwort kennt.

Die entwickelte DeepQA Softwarearchitektur besitzt jedoch noch weitere Funktionalität. Es ist eine Software, um natürliche Sprache sowohl in Fragen als auch in Datenquellen

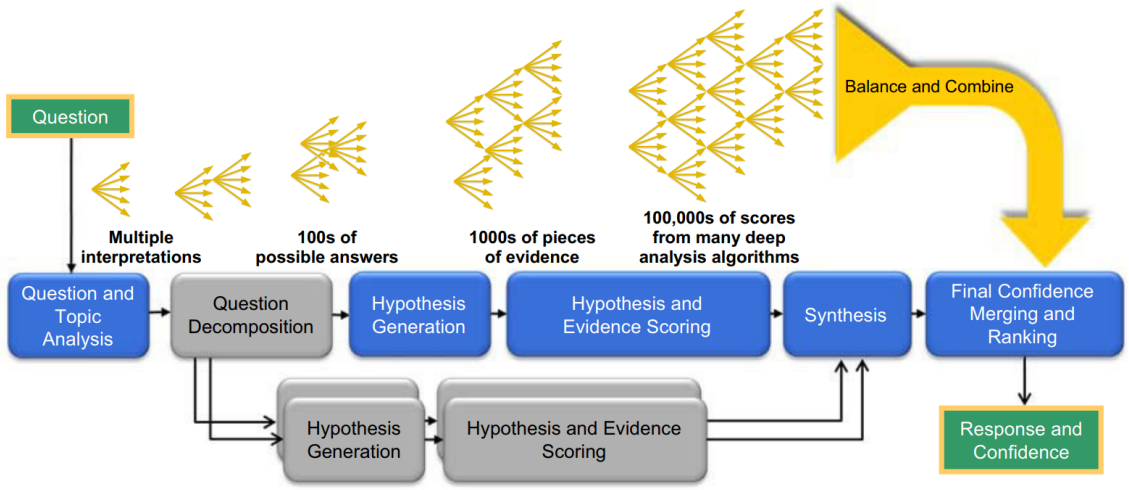

Abb. 2.4: Systemarchitektur von IBM Watson [High 2012, S.4]

verarbeiten zu können. Dabei werden potenzielle Antworten gesammelt und bewertet, indem unstrukturierte Dokumente mit natürlicher Sprache analysiert werden sowie strukturierte Datenquellen wie Datenbanken [High 2012, S.3f; Gliozzo et al. 2013, S.85].

Den Kontext einer Frage zu erkennen ist ein sehr wichtiger Schritt. So musste bei der Beantwortung der Frage bei *Jeopardy!*:„Jodie Foster took this home for her role in 'Silence of the Lambs'“ erkannt werden, dass mit der Phrase „etwas nach Hause bringen“ ein Oscar gemeint war. Die zahlreichen anderen Antwortmöglichkeiten dieser sehr offen gestellten Frage mussten bei herausgefiltert werden [High 2012, S.5].

In Abbildung 2.4 ist die Systemarchitektur von IBM Watson dargestellt. Um eine Frage zu beantworten, wird sowohl die Frage als auch alle möglichen Antworten in der Sammlung des Textkorpus auf hunderte von Arten untersucht, um ein Maß für das Vertrauen in die Interpretation der Frage sowie der möglichen Antwort zu erlangen [High 2012, S.5]:

1. Zuerst wird nach dem Stellen der Frage diese analysiert, um die markantesten Merkmale zu extrahieren.
2. Mehrere Hypothesen werden erstellt, nach Durchsuchung des Textkorpus, bei denen das Potenzial einer möglichen Antwort gegeben ist.
3. Anschließend erfolgt ein tiefer Vergleich der Frage und möglicher Antworten. Hunderte Algorithmen stellen jeweils einen anderen Vergleich an z. B. auf Übereinstimmung von Begriffen und Synonymen sowie Eingrenzungen des Kontexts der Frage.
4. Die Bewertungszahl(en) jedes Algorithmus geben an, inwiefern die potenzielle Antwort aus der Frage abgeleitet werden kann.
5. Alle Punktzahlen werden statistisch gewichtet und ermittelt, wie gut jeder Algorithmus bei der Ermittlung von zwei ähnlichen Schlussfolgerungen in der „Trainingsphase“ war. Hieraus wird ebenfalls das Maß an Vertrauen für die Antwort generiert.
6. Dieser Prozess wird wiederholt für jede der potenziellen Antworten. Dies geschieht so lange, bis Ergebnisse gefunden werden, die bessere Kandidation zur Beantwortung der Frage sind, als die Übrigen.

3 Einsatz von Text Mining in Produktion und Logistik

Im Folgenden wird auf verschiedene Einsatzszenarien von Text Mining für das Themengebiet „Produktion und Logistik“ eingegangen. Ein Beispiel befasst sich mit der Verarbeitung von Kundenkommentaren im Bereich des E-Commerce. Als weiteres Beispiel dient der Produktentstehungsprozess, die damit verbundenen Schulungsaufgaben und wie dabei Text Mining unterstützen kann.

3.1 Retourenvermeidung im E-Commerce

Theoretische Einleitung

Die Ausarbeitung von Walsh und Möhring befasst sich mit der Fragestellung, ob Text Mining dazu beitragen kann, die durch Retouren verursachten Kosten im E-Commerce zu senken. Gleichzeitig soll dabei sichergestellt werden, dass sich die Kundenzufriedenheit nicht negativ verändert [Walsh und Möhring 2014, S.68].

Die Retourenquoten verschiedener Händler variieren im Online-Handel, je nach angebotener Produktgruppe, stark. Die Retouren der Textil- und Bekleidungsbranche weisen Retourenquoten von 50% und mehr auf. Zalando beispielsweise hatte im Jahr 2013 in Deutschland Quoten von ca. 50% der bestellten Paketsendungen, in der Schweiz von ca. 60%. Des Weiteren berichteten 2013 knapp 20% der Online-Händler, dass sie eine steigende Tendenz der Retourenmenge verzeichnen.

Um dieser Entwicklung entgegenzuwirken, werden von den Händlern verschiedene Maßnahmen ergriffen. Amazon ging schon soweit, dass sie Konten von Nutzern mit auffällig hohen Retouren sperrten.

Die durch Retouren verursachten Kosten sind für Unternehmen ein erheblicher Bestandteil im Finanzhaushalt und beeinflussen die Steigerung des Gewinns in hohem Maße negativ. Retouren verursachen zusätzliche Kosten für Transportlogistik, Qualitätsprüfung, Handling, Reinventarisierung sowie den verbundenen Wertverlust der Artikel. Die durchschnittlichen Kosten im Jahr 2013 für eine Retoure lag bei 15,18 € (inkl. Wertverlust) [Walsh und Möhring 2014, S.69f].

Zur Vermeidung von Retouren wird im Folgenden ein Ansatz für präventives Retourenmanagement vorgestellt, welches Retouren vermeiden soll. Diese Maßnahmen sollen sowohl vor als auch nach dem Bestellvorgang des Kunden eingesetzt werden. Es werden unstrukturierte Daten wie z. B. Kundenbewertungen automatisch verarbeitet, um zu Erkenntnissen zu gelangen, die sich für die Vermeidung von Retouren nutzen lassen. In der

Tab. 3.1: Schritte des Text Minings im Retourenmanagement [Walsh und Möhring 2014, S.71]

Prozessschritt	Bezug zum präventiven Retourenmanagement
1. Aufgabendefinition	Erkennung von Retourenmustern in Produktbewertungen
2. Dokumentselektion	Produktbewertungen im eigenen Webshop; Produktbewertungen von Content-Anbietern (evtl. Kundenäußerungen in sozialen Netzwerken)
3. Dokumentaufbereitung	Produktbewertungsaufbereitung durch Zerlegung der Texte in einzelne Wörter/Wortgruppen unter Berücksichtigung der Stammformreduktion (Stemming) und ggf. Löschung von Stoppwörtern und Buchstabentransformationen
4. Text-Mining-Methoden	Gruppierung und Filterung nach retourenrelevanten Wörtern bzw. Wortgruppen
5. Interpretation/Evaluation	Interpretation, ob für das Produkt Retourenmuster vorliegen und ggf. Änderungen nötig sind
6. Anwendung	Implementierung von präventiven Strategien

Vergangenheit war nur eine manuelle Auswertung möglich, was zum einen zeitintensiv ist und sich zum anderen nicht proaktiv einsetzen lässt [Walsh und Möhring 2014, S.70].

Walsh und Möhring gliedern den Prozess der Erkennung von Retourenmustern in sechs Prozessschritte, welche in Tabelle 3.1 dargestellt sind [Walsh und Möhring 2014, S.70f]:

1. In der Aufgabendefinition wird durch die Implementierung der Technologie die Erkennung von Retourenmustern in Produktbewertungen definiert.
2. Anschließend müssen relevante Dokumente selektiert, die untersucht werden sollen. Dies kann sowohl den eigenen Webshop betreffen als auch textuelle Kundenäußerungen in anderen Internetquellen, wie z. B. Bewertungsportale oder soziale Netzwerke.
3. Dieser Schritt dient zur automatisierten Aufbereitung der ausgewählten Dokumente. Dabei werden z. B. die Verfahren Tokenziation, Stemming sowie Stop-Word Removal genutzt.
4. Des Weiteren werden relevante Begriffe für die Erkennung von Retouren definiert. Anschließend wird nach diesen gruppiert und gefiltert (z. B. „fällt größer aus“).
5. Die vorherigen Ergebnisse werden interpretiert und auf die Anzahl überprüft. Eine hohe Anzahl von Wörtern/Wortgruppen kann auf fehlerhafte Produktbeschreibung hinweisen.
6. Der letzte Schritt ist die Ableitung von präventiven Strategien. Das könnte z. B. die Anpassung von Beschreibungen sein oder die Überarbeitung von Produkttexten.

Anwendung am Beispiel von Amazon

Dieser Abschnitt thematisiert die Skizzierung eines möglichen prototypischen Präventionsmanagements zur Reduzierung von Retouren am Beispiel von Amazon. Technische

Details seitens Amazon liegen nicht vor, in welchem Umfang eine derartige Technologie eingesetzt wird.

Zu den häufigsten Gründen für Retouren in der Bekleidungs- und Textilbranche, die Verbraucher angeben, zählen:

- „Artikel gefällt nicht“
- „Artikel passt nicht“
- „mehrere Varianten zur Auswahl bestellt“

Besonders kritisch ist dabei die Abweichung von definierten Konfektionsgrößen unterschiedlicher Hersteller (z. B.: „Schuhgröße 45 fällt bei Produkt X größer aus“). Die Folge dessen ist die Bestellung mehrerer Artikel des Produkts in unterschiedlichen Größen, von denen maximal ein Artikel behalten wird, bei Nichtgefallen keiner [Walsh und Möhring 2014, S.72f].

Zur Bestimmung von Abweichungen innerhalb der Produktbeschreibungen werden im Folgenden die Kundenbewertungen mittels Text Mining untersucht, um Auffälligkeiten zu erkennen. Die auf dem Retourenschein angegebenen Gründe werden zusätzlich dazu genutzt. Beide Datengrundlagen werden nach dem Kauf durch Kunden generiert. Die bereits zuvor geschilderte Vorgehensweise wird nun angewendet, um präventiv Maßnahmen zur Retourenvermeidung umzusetzen [Walsh und Möhring 2014, S.73f].

Die Wahrscheinlichkeit weshalb der Kunde eine Retoure generiert hängt von diversen Faktoren ab. Auf die betreffende Produktgruppe bezogen sind die Gründe aber meistens ähnlich. Eine Untersuchung der häufigsten Retourengründe im Jahr 2013 ergab folgendes Bild [Walsh und Möhring 2014, S.73f]:

- passt nicht
- nicht wie vorgestellt
- kaputt/defekt/beschädigt/mangelhaft
- fällt größer/kleiner aus
- gefällt nicht
- unvollständig
- entspricht nicht der (Produkt-)Beschreibung
- zu spät geliefert
- zu dunkel/zu hell
- zu klein/zu groß

Diese Gründe stellen die Grundlage für die Untersuchung von Kundenbewertungen bei Amazon dar. Walsh und Möhring überprüften jeweils zehn *gut* bis *sehr gut* bewertete Produkte mit zehn *schlecht* bewerteten. Dafür wurden 116 gut bis sehr gut bewertete Jeanshosen mit 28 schlecht bewerteten Damenjeans gegenübergestellt. Die Untersuchung nutzte dabei während des Prozesses relevante Wörter für Retouren sowie Wortgruppen, die die Größe und Farbe definieren. Eingesetzt wurde die Software RapidMiner.

Tab. 3.2: Ergebnisauswertung der Kundenbewertungen mittels Text Mining [Walsh und Möhring 2014, S.75]

	Produktbewertungen		**Durchschnittliche Nennung pro Produktbewertung**	
	Schlecht bewertete Jeans (n=28)	Gut bewertete Jeans (n=116)	Schlecht bewertete Jeans	Gut bewertete Jeans
Fällt größer aus (bspw. groß, viel zu groß, groß geschnitten)	24	17	0,85	0,14
Fällt kleiner aus (bspw. kleiner, fällt kleiner aus, kleiner bestellen)	23	30	0,82	0,25
Hellere Farbe (bspw. heller, hellere Variante)	6	7	0,21	0,06
Dunklere Farbe (bspw. dunklere Abbildung, dunkler)	8	17	0,28	0,14

Die Ergebnisse der Auswertung sind in Tabelle 3.2 dargestellt. Nach Durchführung der in Tabelle 3.1 beschriebenen Schritte 1 bis 4 fällt auf, dass retourenrelevante Wörter häufiger bei negativ bewerteten Produkten auftreten, als bei gut bewerteten: Durchschnittliche Nennung „fällt größer aus“ bei schlecht bewerteten Jeans (0,85 Nennungen) ist signifikant höher im Vergleich zu gut bewerteten Jeans (0,14 Nennungen).

Die Schlussfolgerung aus der Auswertung ist die Ableitung der Maßnahme, dass bei Überschreitung eines definierten Schwellenwerts bestimmter Wörter/Wortgruppen eingegriffen werden muss, um nicht zu viele Retouren der Kunden zu erhalten. Der Schwellenwert muss dabei so gewählt werden, dass beispielsweise scherzhafte und ironische Bewertungen noch keine Aktion auslösen.

Bei betreffenden Produkten sollte anschließend automatisiert ein Hinweis auf der Seite des Produkts erscheinen z. B. „Produkt fällt größer aus“, damit Kunden diese Information vor der Bestellung bereits berücksichtigen können. Ein weiterer Hinweis wäre vor Abschluss des Bestellvorgangs im Warenkorb sinnvoll.

Sollte bei einzelnen Produkten eine auffällig hohe Nennung von kritischen Begriffen erfolgen, sollten diese zunächst für die Bestellung gesperrt werden und einer manuellen Überprüfung unterzogen werden. Dieser Vorgang würde vor Fehllieferungen bei Material- oder Herstellungsfehlern bestimmter Produkte schützen. Des Weiteren sollte diese Informationen innerhalb der Supply Chain an die Lieferanten und Hersteller weitergereicht werden, um die Anzahl überflüssiger Logistikprozesse zu minimieren.

Die beschriebenen Maßnahmen führen dazu, dass die Zahl der Rücksendungen gesenkt werden können, ohne dabei das Kauferlebnis der Kunden negativ zu beeinflussen. Das Erlebnis kann sogar positiv verändert werden, da kein Umtausch von Artikeln nötig ist und ungewollte Retouren die Kundenzufriedenheit senken.

3.2 Moderne Produktentstehungsprozesse: Erfassung von Simulationswissen

Breitsprecher et al. beschreiben im Folgenden die Anwendung von Text Mining, um bereits gewonnenes Wissen aus Simulationsergebnissen und Berechnungsberichten weiterhin Mitarbeiten zur Verfügung stellen zu können. Der Fokus wird dabei auf die Effizienzsteigerung der Produkt- und Prozessentwicklung gelegt. Das Ziel ist die Erstellung eines wissensbasierten FEA-Assistenzsystems, das weniger erfahrene Simulationsanwender bei der Erstellung und Auswertung von FEA-Analysen unterstützt. Die genutzten Daten sind bereits validierte Simulationsmodelle aus Berichten, Informationen über geometrische Vereinfachungen sowie Kontakt- und Randbedingungen [Breitsprecher et al. 2015, S.744f].

Zunächst müssen für das geplante Vorhaben alle Informationen der notwendigen Datenbestände zusammengestellt werden. Die Überführung von Simulationsparametern und Einstellungen aus validierten Modellen in Tabellen ist noch mit vergleichsweise überschaubarem Aufwand durchzuführen. Bei einer Vielzahl von unstrukturierten Daten (z. B. Berechnungsberichte) ist der Arbeitsaufwand erheblich höher. Diese Daten müssen zunächst aufbereitet und strukturiert werden. Dies geschieht mit Methoden aus dem Bereich des Text Minings. Vor allem nutzen Breitsprecher et al. für diesen Prozess Textklassifikation und Informationsextraktion.

Durch die Textklassifikation werden Berechnungsberichte nach Klassen (z. B. FE-Analyseart) der Ergebnisgröße oder der betrachteten Bauteile gruppiert. Für diesen Vorgang werden die Häufigkeiten der Wörter (Token) aus den Berichten in einer Term-Dokumenten-Matrix dargestellt. Eine solche Darstellung ist in Abbildung 3.1 abgebildet. Diese Form eignet sich zum Auffinden von Schlüsselwörtern, welche für den jeweiligen Bericht von besonders hoher Wichtigkeit sind. Breitsprecher et al. untersuchten, dass der Begriff „Kraft“ wesentlich häufiger in Berichten aus statischen Analysen zu finden ist, als aus Modalanalysen. Der Begriff „Frequenz“ verhält sich genau entgegengesetzt [Breitsprecher et al. 2015, S.747f].

Betreffende Begriffe werden mittels Stemming auf den Wortstamm zurückgeführt, um deren Signifikanz zu steigern. In siehe Abbildung 3.1 wird dies für die Attribute 2 und 5 durchgeführt. Mittels Klassifikationsmethoden (z.B k-Nearest-Neighbour) werden die Berichte schließlich anhand ihrer Ähnlichkeit gruppiert. Dazu werden die Häufigkeiten der Term-Dokumenten-Matrix genutzt.

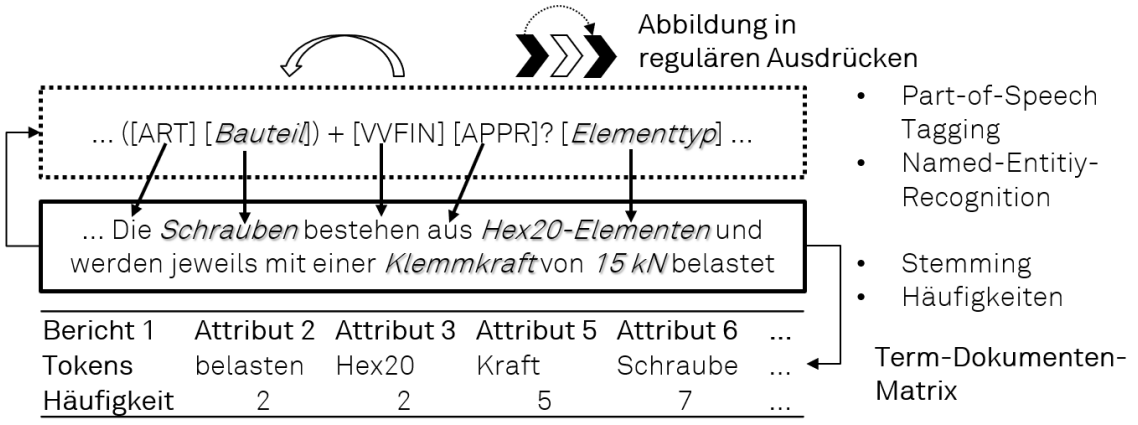

Abb. 3.1: Analyse unstrukturierter Berechnungsberichte durch Text Mining [Breitsprecher et al. 2015, S.747]

Nach der Klassifikation aller Berichte können nun relevante Textauszüge ausgegeben werden. Um die angefragte Information zu liefern, werden Kategorien und Suchbegriffe verwendet. Die Aufbereitung dafür erfolgt durch Informationsextraktion. Mit dieser Methode lassen sich die relevanten Bestandteile der Berichte extrahieren [Breitsprecher et al. 2015, S.747].

Mittels POS Tagging werden den einzelnen Wörtern ihre jeweiligen Wortarten im Satz zugeordnet. In Abbildung 3.1 sind die Artikel „ART“, Verben „VVFIN“ und Präpositionen „APPR“ dargestellt. Für diesen Vorgang wird ein Tagset verwendet wie z. B. das bekannte Stuttgart-Tübingen-Tagset. Weiterhin werden mittels Named-Entitiy-Recognition Oberbegriffsklassen gebildet, diese sind in der Abbildung durch die zugeordneten Wortklassen „Elementtyp“ und „Bauteil“ der Substantive/Eigennamen ersichtlich.

Im letzten Schritt werden betreffende Textausschnitte mit regulären Ausdrücken abgebildet, um weitere Passagen automatisch erfassen zu können. Die Ausdrücke werden dabei wie Schablonen genutzt, die sich über wichtige Textpassagen legen lassen können. Dargestellt sind in Abbildung 3.1 die Operatoren regulärer Ausdrücke, um Wiederholungen (+) oder optionale Textbestandteile zu definieren (?). Der Schritt der Zuweisung von Wortarten und Oberbegriffsklassen dient dazu, den Ausdrücken die notwendige Flexibilität zu geben, um diese auf andere Texte anwenden zu können [Breitsprecher et al. 2015, S.748].

Die beschriebene Anwendung von Text Mining führt zu einem Prozess, welcher automatisiert erforderliches Simulationswissen erfasst. Die Informationen werden dabei aus unstrukturierten Berechnungsberichten und validierten Modellen gewonnen. Dieser Vorgang kann parallel zu manuellen Akquisitionsmethoden genutzt werden, um somit eine fundierte Basis an Wissen für das FEA-Assistenssystem zu erhalten. Abschließend kann so sichergestellt werden, dass für Simulationsaufgaben brauchbare Simulationsmodelle automatisiert generiert werden können [Breitsprecher et al. 2015, S.748].

4 Zusammenfassung und Ausblick

Die vorliegende Arbeit hatte als Ziel Text Mining und dessen Verfahren aufzubereiten. Durch die zunehmende Nutzung des Internets und der steigenden Informationsmenge in Form von Textdokumenten gewinnt Text Minig zunehmend an Bedeutung. Für das Wissensmanagement, besonders für Unternehmen, ist es ein sehr wichtiger Baustein, um beschäftigen Personen den Zugang zu Unternehmenswissen zu ermöglichen. Dieses Wissen ist entscheidend für den zukünftigen Erfolg des Unternehmens. Weiterhin lassen sich genauere Analysen erstellen und es kann besser auf die Anforderungen der Kunden eingegangen werden. Dabei können überflüssige Prozesse eingespart werden, die sowohl Kapital einsparen, als auch der Umwelt zu Gute kommen.

Im Themenumfeld der Logistik scheint die Thematik noch nicht flächendeckend eingesetzt werden und es ist noch Potenzial für die Implementierung vorhanden. Die Recherche nach Beispielen aus der Logistik erwies sich als schwierig. Dies kann eventuell aber auch an der Informationspolitik der Unternehmen liegen, dass derartige IT-Strukturen nicht detailliert und gut dokumentiert nach außen getragen werden. Ebenfalls auffällig war, dass größtenteils englischsprachige Literatur zum Thema Text Mining existiert.

Zusammenfassend ist festzuhalten, dass es sich bei Text Mining aufgrund der sehr vielfältigen Anwendungsgebiete um ein sehr großes und dynamisches Gebiet der Forschung handelt. In der Zukunft wird sich die Anwendung vermehrt auf das Internet abzeichnen; im Zuge der zunehmenden Vernetzung. Die Bereitstellung von Wissen wird auch zukünftig immer wichtiger, sei es durch das bloße Verfügbarmachen oder die Generierung von neuem Wissen, aus bereits vorhandenen Daten. Aus diesem Grund ist Text Mining eine sinnvolle Möglichkeit, um die Datenflut kontrollierbar zu halten.

Literatur

Breitsprecher, Thilo, Philipp Kestel, Christof Küster, Tobias Sprügel und Sandro Wartzack (Nov. 2015). „Einsatz von Data-Mining in modernen Produktentstehungsprozessen: Ganzheitliche Forschung für Ingenieure von morgen“. In: *ZWF Zeitschrift für wirtschaftlichen Fabrikbetrieb* 110, S. 744–750. DOI: 10.3139/104.111423.

Christopher Manning, Prabhakar Raghavan und Hinrich Schuetze (2009). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press. ISBN: 9780521865715.

Feldman, Ronen und Ido Dagan (1995). „Knowledge Discovery in Textual Databases (KDT)“. In: *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*. KDD'95. Montreal, Quebec, Canada: AAAI Press, S. 112–117. URL: http://dl.acm.org/citation.cfm?id=3001335.3001354.

Feldman, Ronen und James Sanger (2008). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Reprinted. Cambridge: Cambridge Univ. Press. ISBN: 978-0-521-83657-9.

Gliozzo, Alfio, Or Biran, Siddharth Patwardhan und Kathleen McKeown (Aug. 2013). „Semantic Technologies in IBM Watson“. In: *Proceedings of the Fourth Workshop on Teaching NLP and CL*. Sofia, Bulgaria: Association for Computational Linguistics, S. 85–92. URL: https://www.aclweb.org/anthology/W13-3413.

Hearst, Marti A. (1999). „Untangling Text Data Mining“. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. ACL '99. College Park, Maryland: Association for Computational Linguistics, S. 3–10. ISBN: 1-55860-609-3. DOI: 10.3115/1034678.1034679.

Heyer, Gerhard, Uwe Quasthoff und Thomas Wittig (2012). *Text Mining: Wissensrohstoff Text (Konzepte, Algorithmen, Ergebnisse)*. Korrigierter Nachdr. Herdecke: W3L-Verl. ISBN: 3-937137-30-0.

High, Rob (2012). „The era of cognitive systems: An inside look at IBM Watson and how it works“. In: *Redbooks (REDP-4955-00), IBM Corporation*.

Hotho, Andreas, Andreas Nürnberger und Gerhard Paaß (Mai 2005). „A Brief Survey of Text Mining“. In: *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology* 20.1, S. 19–62. ISSN: 0175-1336.

Jo, Taeho (2019). *Text Mining*. Bd. 45. Cham: Springer International Publishing. ISBN: 978-3-319-91814-3. DOI: 10.1007/978-3-319-91815-0.

Kosala, Raymond und Hendrik Blockeel (Juni 2000). „Web Mining Research: A Survey“. In: *SIGKDD Explor. Newsl.* 2.1, S. 1–15. ISSN: 1931-0145. DOI: 10.1145/360402.360406.

Lanquillon, Carsten (2001). „Enhancing Text Classification to Improve Information Filtering“. Diss. Otto-von-Guericke-Universität, Magdeburg.

Mehler, Alexander und Christian Wolff (2005). „Einleitung: Perspektiven und Positionen des Text Mining“. In: *LDV-Forum* 20.1, S. 1–18. URL: https://epub.uni-regensburg.de/6844/.

Pfeifer, Katja (2014). „Serviceorientiertes Text Mining am Beispiel von Entitätsextrahierenden Diensten“. Diss. Technische Universität Dresden.

Reinsel, David, John Gantz und John Rydning (2018). *The Digitization of the World: From Edge to Core*. URL: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.

Rijsbergen, C. J. Van (1979). *Information Retrieval*. 2nd. Newton, MA, USA: Butterworth-Heinemann. ISBN: 0408709294.

Sebastiani, Fabrizio (März 2002). „Machine Learning in Automated Text Categorization“. In: *ACM Comput. Surv.* 34.1, S. 1–47. ISSN: 0360-0300. DOI: 10.1145/505282.505283.

Walsh, Gianfranco und Michael Möhring (Feb. 2014). „Retourenvermeidung im E-Commerce – Kann Big Data helfen?“ In: *Marketing Review St. Gallen* 31.1, S. 68–78. DOI: 10.1365/s11621-014-0322-6.

Weiss, Sholom M., Nitin Indurkhya und Tong Zhang (2015). *Fundamentals of Predictive Text Mining*. London: Springer London. ISBN: 978-1-4471-6749-5. DOI: 10.1007/978-1-4471-6750-1.

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

HTML	Hypertext Markup Language
PDF	Portable Document Format
POS	Part-of-Speech
STTS	Stuttgart-Tübingen-Tag-Set
TF	Term Frequency
TF-IDF	Term Frequency-Inverse Term Frequency
XML	Extensible Markup Language