Data Mining und Big Data Analytics: Semantische Suche, Prognose und Entscheidungsunterstützung mit Künstlichen Neuronalen Netzen

Der Wirtschaftswissenschaftlichen Fakultät der
Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des akademischen Grades

Doktor der Wirtschaftswissenschaften
- Doktor rerum politicarum –

vorgelegte Dissertation

von

Diplom-Ökonom Christoph Gleue

[redacted]

2018

Betreuer und Gutachter:	Prof. Dr. Michael H. Breitner
Weiterer Gutachter:	Prof. Dr. J.-Matthias Graf von der Schulenburg
Vorsitzender der Prüfungskommission:	Prof. Dr. Stefan Wielenberg
Weiteres Mitglied (beratend):	Dr. Michelle Muraz

I. Abstract

This cumulative dissertation summarizes and critically reflects on a total of six scientific publications in which I was involved as an author. The overall objective of this thesis is the analysis and discussion of the question of how traditional, historically grown workflows and business processes can be optimized using methods and technologies from the broad fields of data mining, big data analytics, and data science. It is of particular interest how this can be accomplished successfully, beyond the large "lighthouse projects" in industry and research but in small and medium-sized projects in science and in the private sector. All publications are based on two independent research projects:

(1) Semantic Similarity Analysis in the Literature Research Process: This section presents the development, demonstration and benchmarking of a novel tool for semantic literature research based on Latent Semantic Indexing (LSI). The "Tool for Semantic Indexing and Similarity Queries" (TSISQ) enables scientists, regardless of their field of research, to quickly and easily identify semantically similar research papers. In addition, the structure and architecture of the system are presented, hints for its practical application are given, and the efficiency and quality improvements that can be achieved through the use of TSISQ in the literature research process are pointed out.

(2) Decision Support for the automotive industry: This section describes the results and findings of a cooperation project with a large German car manufacturer, in which a comprehensive decision support system was designed and implemented. The most important core functionality of this system is the forecast of market and residual values for used vehicles using Artificial Neural Networks. In addition to the presentation of the findings in the field of long-term forecasts, our results regarding internal (influenceable by the manufacturer) and external (macroeconomic) influencing factors on vehicle residual values are shown and critically discussed.

Keywords: Semantic Search, Latent Semantic Indexing, Information Retrieval, Text Mining, Literature Research, Decision Support, Data Mining, Big Data Analytics, Data Warehouse, Artificial Neural Networks, Leasing, Residual Value Management, Used Vehicles

Kurzdarstellung

Diese kumulative Dissertation ist eine Zusammenfassung und kritische Reflexion von insgesamt sechs wissenschaftlichen Publikationen, an denen ich als Autor beteiligt war. Das übergeordnete Ziel dieser Arbeit ist die Analyse und Diskussion der Frage, wie abseits der großen „Leuchtturmprojekte“ in Industrie und Forschung auch in kleineren und mittleren Projekten in Wissenschaft und Privatwirtschaft mit Methoden und Technologien aus dem weiten Feld des Data Mining, Big Data Analytics und Data Science althergebrachte Arbeitsabläufe und Geschäftsprozesse erfolgreich optimiert werden können. Alle Veröffentlichungen basieren auf zwei voneinander unabhängigen Forschungsprojekten:

(1) Semantische Ähnlichkeitsanalyse im Literaturrechercheprozess: In diesem Abschnitt werden Entwicklung, Demonstration und Benchmark eines neuartigen Werkzeugs zur semantischen Literaturrecherche basierend auf Latent Semantic Indexing (LSI) vorgestellt. Das „Tool for Semantic Indexing and Similarity Queries“ (TSISQ) ermöglicht Wissenschaftlern, unabhängig von ihrem Forschungsgebiet, die schnelle und einfache Identifikation von semantisch ähnlichen Forschungsarbeiten. Es werden zudem Aufbau und Architektur des Systems präsentiert, Hinweise für dessen praktische Anwendung gegeben und die Effizienz- und Qualitätsverbesserungen aufgezeigt, die durch den Einsatz von TSISQ im Literaturrechercheprozess erzielt werden können.

(2) Entscheidungsunterstützung für die Automobilindustrie: Dieser Abschnitt beschreibt die Ergebnisse und Erkenntnisse eines Kooperationsprojektes mit einem großen deutschen Automobilhersteller, wobei ein umfangreiches Entscheidungsunterstützungssystem (EUS) konzipiert und implementiert wurde. Die wichtigste Kernfunktionalität dieses Systems stellt dabei die bedarfsgerechte Prognose von Markt- und Restwerten für Gebrauchtfahrzeuge mit Künstlichen Neuronalen Netzen (KNN) dar. Neben der Vorstellung der Erkenntnisse im Bereich der Langfristprognosen werden auch unsere Ergebnisse bzgl. interner (vom Hersteller beeinflussbarer) und externer (makroökonomischer) Einflussfaktoren auf Fahrzeugrestwerte präsentiert.

Schlagworte: Semantische Suche, Latent Semantic Indexing, Information Retrieval, Text Mining, Literaturrecherche, Entscheidungsunterstützung, Data Mining, Big Data Analytics, Data Warehouse, Künstliche Neuronale Netze, Leasing, Restwertmanagement, Gebrauchtfahrzeuge

II. Management Summary

Problemstellung und Forschungsziele

Data Mining, Business Intelligence, Big Data Analytics und Data Science gehören seit einigen Jahren in nahezu allen Branchen zu den wichtigsten Zukunftsthemen. Die Nachfrage nach Mitarbeitern mit entsprechenden Fähigkeiten ist inzwischen deutlich höher als das Angebot (vgl. Davenport und Patil, 2012). Auch in der Wirtschaftsinformatik als wissenschaftlicher Disziplin steigt seit Jahren stetig die Anzahl an Publikationen in diesem überaus umfangreichen Themengebiet. Moderne Technologien, benutzerfreundlichere Analysesoftware, ein hoher Vernetzungsgrad und vergleichsweise günstige Rechen- und Speicherkapazitäten machen die Verarbeitung und Analyse von „Big Data“ inzwischen einem breiteren Publikum zugängig (vgl. Agarwal und Dhar, 2014).

Ziel dieser Dissertation ist es, zu untersuchen, wie abseits der großen „Leuchtturmprojekte“ in Industrie und Forschung mit Methoden und Technologien aus dem weiten Feld des Data Mining, (Big) Data Analytics und Data Science althergebrachte Arbeitsabläufe automatisiert und optimiert werden können. Dazu werden im Folgenden zwei Projekte mit jeweils drei Veröffentlichungen vorgestellt, die exemplarisch zeigen sollen, wie auch kleinere und mittlere Projekte in der Welt der Wissenschaft und der Privatwirtschaft erfolgreich durchgeführt werden können und welche Herausforderungen sich dabei typischerweise ergeben. Diese kumulative Doktorarbeit besteht aus zwei unabhängigen Teilen. Beide Teile bewegen sich im Themenkomplex des Data Mining und Big Data Analytics, basieren jedoch auf zwei voneinander unabhängigen Projekten.

Semantische Ähnlichkeitsanalyse im Literaturrechercheprozess:

In Kapitel 2 werden die Erkenntnisse aus drei Forschungsarbeiten zusammengefasst, die einen Beitrag zum Themengebiet des Text Mining bzw. der semantischen Textanalyse zur Unterstützung der Literaturarbeit in der wissenschaftlichen Praxis leisten. Als Ergebnis ist dabei ein Werkzeug zur automatisierten Literaturrecherche auf Basis der Erkennung semantischer Ähnlichkeiten zwischen Texten entstanden, das „Tool for Semantic Indexing and Similarity Queries“ (TSISQ). Die folgenden Themen und Publikationen bauen aufeinander auf und werden im ersten Hauptteil der vorliegenden Doktorarbeit adressiert:

- Koukal A., Gleue, C. und Breitner, M. H. (2013). „Enhancing Literature Research Processes: A Glance at an Approach Based on Latent Semantic Indexing“, veröffentlicht in den Proceedings der Gesellschaft für Informatik (GI) 2013, „Lecture Notes in Informatics (LNI)”,
- Koukal A., Gleue, C. und Breitner, M. H. (2014). „Enhancing Literature Review Methods - Towards more efficient Literature Research with Latent Semantic Indexing“, veröffentlicht in den Proceedings der 22. European Conference on Information Systems (ECIS 2014),
- Koukal A., Gleue, C. und Breitner, M. H. (2014). „Enhancing Literature Review Methods: Evaluation of a Literature Search Approach based on Latent Semantic Indexing“, veröffentlicht in den Proceedings der International Conference on Information Systems (ICIS 2014).

Die übergeordnete Motivation des in Kapitel 2 vorgestellten Forschungsprojekts ist es, den Literaturrechercheprozess als eine der zeitintensivsten, wiederkehrenden Aufgaben für Wissenschaftler jeden Fachgebiets mithilfe von Methoden aus dem Gebiet des Text Mining und der semantischen Suche zu optimieren (siehe Abbildung 1). Die dabei entstehende Zeitersparnis kann somit für andere Forschungstätigkeiten aufgewendet werden und, neben einer Vervollständigung des theoretischen Fundaments, insgesamt eine Qualitätsverbesserung der Forschungsarbeiten bei gleichbleibendem Zeitaufwand ermöglichen.

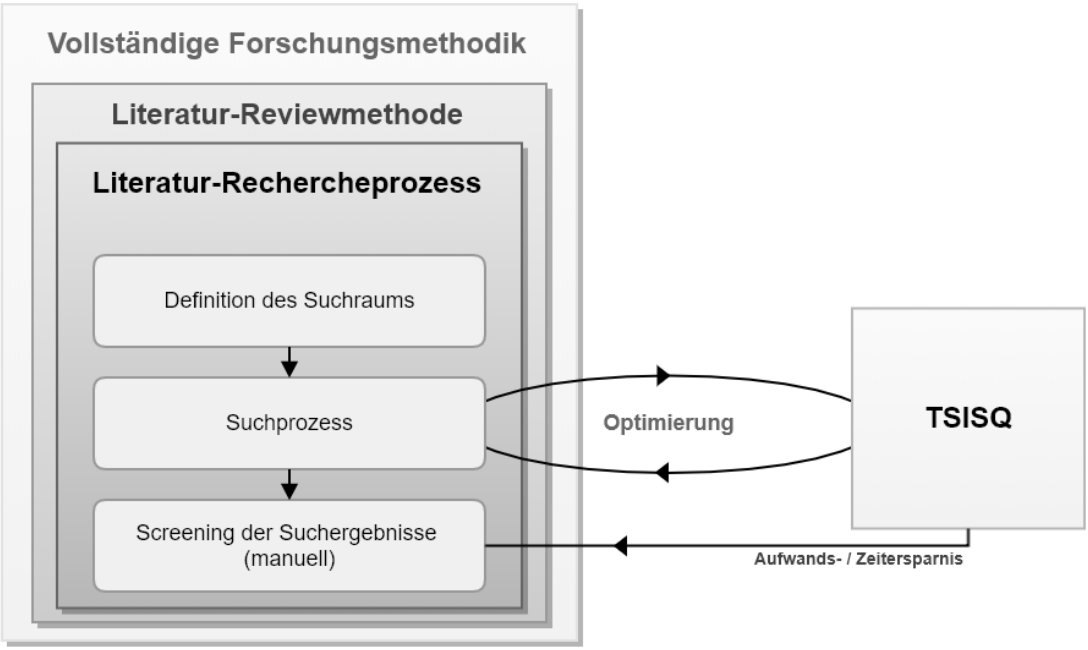

Abbildung 1: Optimierung des Literaturrechercheprozesses mit TSISQ

Die fehlende Automatisierung von Volltextsuchen in Kombination mit einer mangelnden Berücksichtigung semantischer Konzepte bzw. semantischer Ähnlichkeiten trägt wesentlich dazu bei, dass die Literaturrecherche im Rahmen des wissenschaftlichen Arbeitens ein langsamer, zeitintensiver (weil größtenteils manueller) Prozess ist. Um hier eine optimale technologische Unterstützung zu gewährleisten, bedarf es eines effizienteren, toolgestützten Ansatzes, der den manuellen Anteil an der Literaturrecherche, beispielsweise beim Screening der initialen Suchergebnisse, auf ein Minimum reduziert. Diese Idee führte letztlich zu der Entwicklung von TSISQ.

Stichwortbasierte Suchverfahren bringen trotz ihrer einfachen Realisierbarkeit Nachteile mit sich (vgl. Blair und Maron, 1985; Homayouni et al., 2004; LaBrie und St. Louis, 2003) und sind daher alles andere als optimal (vgl. Dumais et al., 1988): Ambiguität von Suchbegriffen, Synonymie, Polysemie, die unsachgemäße Anwendung von Stoppwörtern wie „die“, „in“ oder „und“, Bindestriche, Klammern, Pluralformen sowie letztlich die auftretende Unschärfe und Inkonsistenz bei der Indizierung des zu durchsuchenden Textmaterials führen schnell zu einer Verzerrung der Suchergebnisse. Folglich führen stichwortbasierte Ansätze sehr wahrscheinlich zu falsch-positiven oder falsch-negativen Suchergebnissen. Anders ausgedrückt werden potenzielle Treffer nicht gefunden oder schlicht unzutreffende Suchergebnisse geliefert (vgl. Blair und Maron, 1985; Dumais et al., 1988; Hofmann, 1999; LaBrie und St. Louis, 2003; Salton und McGill, 1986; Yandell und Majoros, 2002).

Trotz dieser Defizite sind rein stichwortbasierte Suchverfahren auch heute noch bei einem Großteil der wissenschaftlichen Datenbanken üblich (vgl. Cui et al., 2003). Daher wurden alle für die Wirtschaftsinformatik relevanten wissenschaftlichen Datenbanken wie AISel, IEEE, JStor, ScienceDirect oder Wiley untersucht und dabei festgestellt, dass sich das beschriebene Fehlerverhalten in den Suchergebnissen widerspiegelt, was darauf schließen lässt, dass eine Reduzierung oder gar Beseitigung der zuvor genannten Defizite stichwortbasierter Suchverfahren einen wichtigen Beitrag darstellt. Um einer Lösung für die hier umrissene Problemstellung näherzukommen ist es das Ziel unserer Forschung zu diesem Thema, einen alternativen, modernen und intelligenteren Ansatz zur automatisierten Ähnlichkeitsanalyse von Texten zu entwickeln, der die folgenden Kriterien erfüllt:

- Verlässliche Identifikation wissenschaftlicher Arbeiten, die einem bestimmten Forschungsfeld angehören (Vorhandensein der initialen Suchbegriffe).

- Identifikation semantisch ähnlicher Arbeiten aus demselben oder sogar aus anderen Forschungsgebieten (Vorhandensein von Synonymen und/oder mit den initialen Suchbegriffen „sinnhaft verwandter“, also semantisch ähnlicher Begriffe).
- Die Überwindung des Problems mit Synonymie und Polysemie von Suchbegriffen und somit die größtmögliche Vermeidung falsch-positiver und falsch-negativer Suchergebnisse.
- Die Möglichkeit, Suchanfragen in natürlicher Sprache zu formulieren (Verwendung von Stichwörtern, vollständigen Sätzen, Abstracts und ganzen Forschungsarbeiten als Suchanfrage).

Das zu diesem Zweck entwickelte System ist schematisch in Abbildung 2 dargestellt. Zunächst wird dazu eine beliebig große Zahl wissenschaftlicher Arbeiten im PDF- oder Textformat eingelesen. Stark vereinfacht ausgedrückt wird anschließend mithilfe eines Verfahrens aus dem Fachgebiet des Information Retrieval, „Latent Semantic Indexing“ (LSI), eine numerische Repräsentation der Volltexte als Vektoren in einem Vektorraum erzeugt und einem Gesamtindex hinzugefügt, der als Basis für zukünftige Suchanfragen dient. Nach erfolgtem Training des Gesamtindex können einzelne Suchanfragen auf die gleiche Weise transformiert und die resultierenden Vektoren mit denen innerhalb des Index verglichen werden.

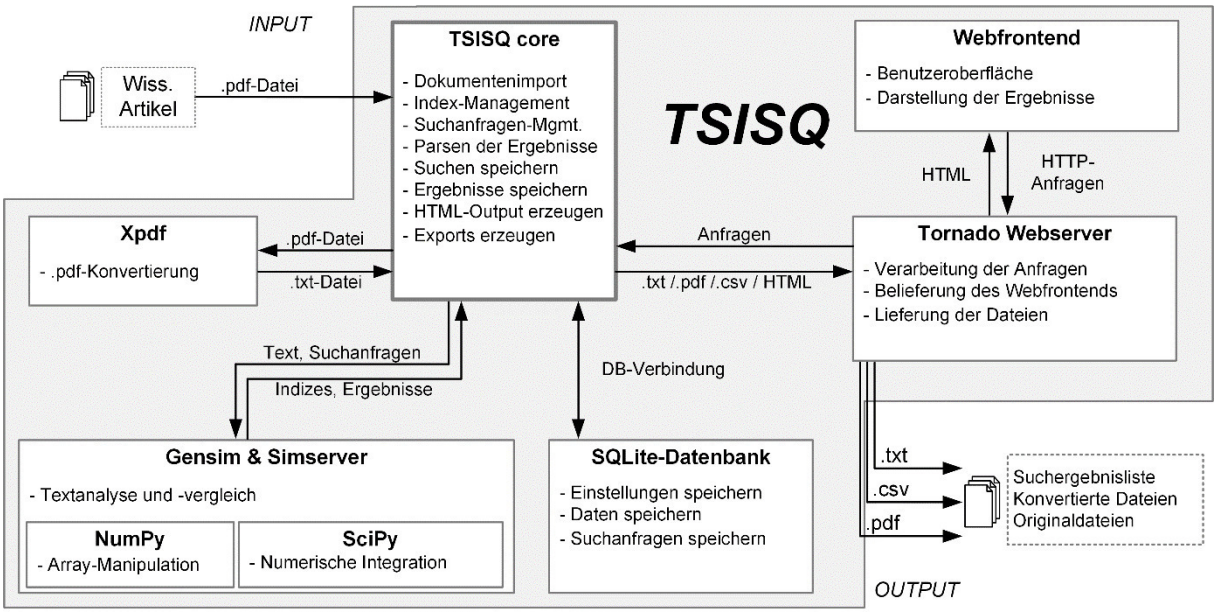

Abbildung 2: Tool for Semantic Indexing and Similary Queries

Abhängig von dem Winkel, in dem die Vektoren letztlich zueinander stehen, kann eine Aussage über die semantische Ähnlichkeit der durch sie repräsentierten Texte getroffen werden: Je größer der Winkel ist, desto größer ist auch die semantische Abweichung. Suchanfragen an TSISQ können in natürlicher Sprache über das dazugehörige Webfrontend gestellt werden,

sowohl als Upload im PDF- oder TXT-Format als auch per Volltexteingabe in ein Textfeld. Sobald der Gesamtindex einmal trainiert ist erscheint bei Klick auf den „Start Query“-Button (vgl. Abbildung 13) ohne nennenswerte Zeitverzögerung die Suchergebnisliste mit den als semantisch ähnlich identifizierten Artikeln, in absteigender Reihenfolge bzgl. ihrer Ähnlichkeit zur Suchanfrage (vgl. Abbildung 14). Mit dem Ziel, Eignung und Nutzen von TSISQ für den Literaturrechercheprozess zu bewerten und messbar zu machen, wurden in mehreren Anwendungsfällen zweierlei Tests durchgeführt:

1. Ein Benchmark zu den Ergebnissen einer stichwortbasierten Suchmaschine (mit Unterstützung von Fachexperten) auf dem 12.332 Paper umfassenden Index aus den führenden Zeitschriften in der Wirtschaftsinformatik, des sog. „AIS Basket of Eight“ und den renommiertesten Konferenzen. Auf diesem Wege wurden ein Vergleich und eine Bewertung der falsch-positiven Fehlerraten der verschiedenen Suchverfahren durchgeführt.
2. Ein Laborexperiment in einer kontrollierten Umgebung, das auf eine Bemessung der falsch-positiven und (insbesondere) der falsch-negativen Fehlerrate abzielt.

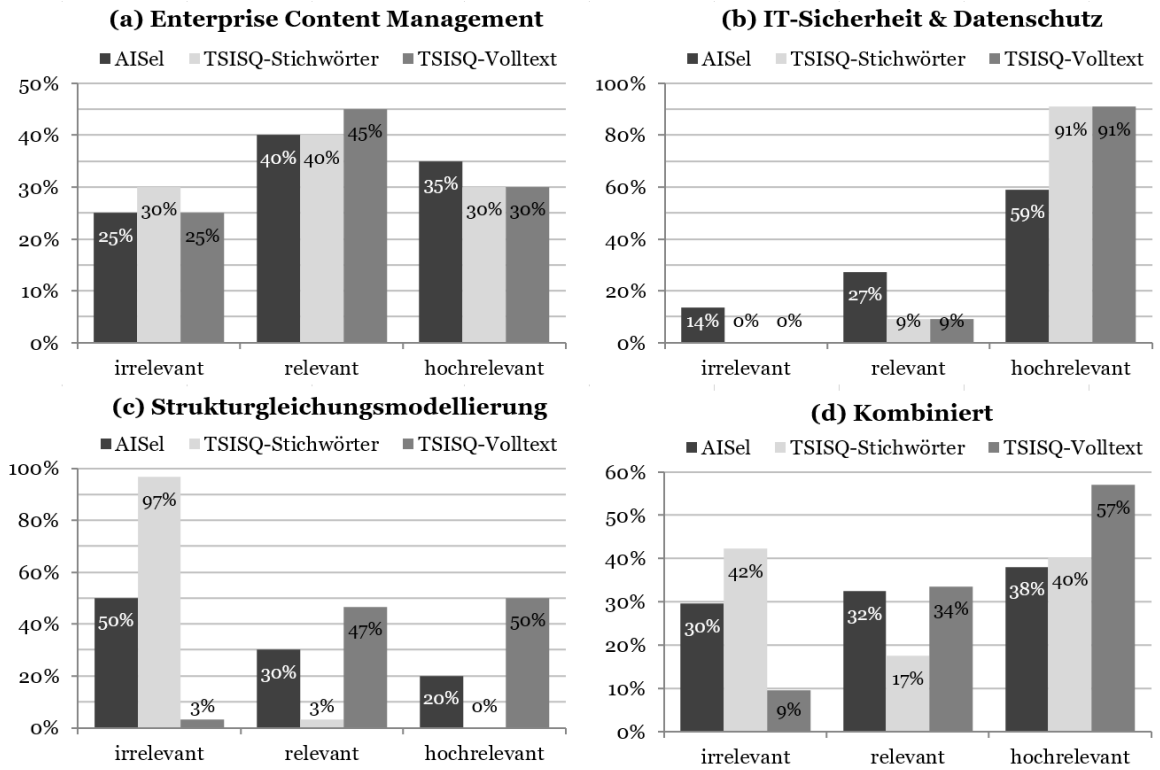

Abbildung 3: Ergebnisse des Benchmarks von TSISQ gegenüber Schlagwortsuchen

Zu 1.: Der Benchmark bezieht sich auf eine durch Fachexperten in den jeweiligen Gebieten „Enterprise Content Management“ (ECM), „IT-Sicherheit und Datenschutz“ und „Strukturgleichungsmodellierung“ (SEM) durchgeführte Bewertung der Relevanz der

Suchergebnisse auf dem großen Gesamtindex in „irrelevant“, „relevant“ und „hoch relevant“. Es wurden zudem geeignete Suchanfragen formuliert, um die den entsprechenden Fachgebieten zugehörigen Artikel im Gesamtindex zu identifizieren.

Zu 2.: Das Laborexperiment zielt darauf ab, die falsch-positive und insbesondere die falsch-negative Fehlerrate von TSISQ zu bemessen. Es soll demonstriert werden, wie viele den o.g. Themengebieten zugehörige Publikationen durch TSISQ „übersehen“ werden und wie viele Suchergebnisse fälschlicherweise als dazu passend identifiziert wurden. Dazu wurde ein deutlich kleinerer, kontrollierbarer Index aus 100 zufällig ausgewählten Veröffentlichungen manuell erstellt und anschließend deren Inhalte gemäß der unter (i) vorgestellten Kategorisierung für das Thema „IT-Sicherheit und Datenschutz“ bewertet. TSISQ lieferte ein Suchergebnis von insgesamt 23 Ergebnissen, obwohl im Gesamtindex nur 18 Artikel von 100 zur Zieldomäne gehörten. Die 23 gefundenen Ergebnisse enthielten dabei 15 (mindestens) relevante und 8 irrelevante Artikel. Folglich wurden 3 der 18 insgesamt relevanten Artikel nicht identifiziert, was einer falsch-negativen Fehlerquote von 17 Prozent entspricht. Dabei waren die ersten fünf Suchergebnisse mit einer Trefferquote von 100 Prozent am genauesten. Während im Bereich der Top 17-Ergebnisse die durchschnittliche falsch-positive Fehlerrate mit weniger als 20 Prozent immer noch auf einem akzeptablen Niveau liegt, erhöht sich diese Rate auf 35 Prozent, wenn alle 23 Suchergebnisse berücksichtigt werden (vgl. Abbildung 4).

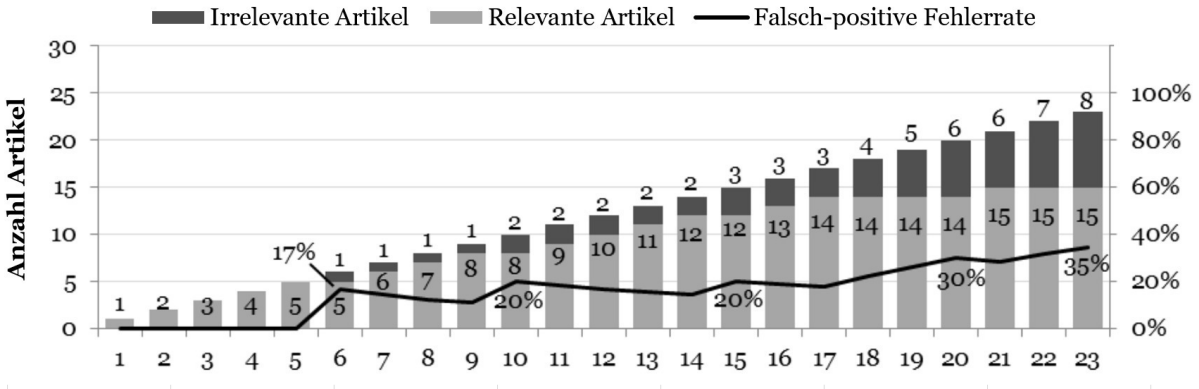

Abbildung 4: Ergebnisse des Laborexperiments

Das Laborexperiment erlaubt eine zusätzliche Quantifizierung der falsch-positiven und falsch-negativen Fehlerrate, deren Reduktion eine weitere der zu Beginn geforderten Eigenschaften darstellt. In dieser kontrollierten Umgebung beträgt die durchschnittliche Suchgenauigkeit mindestens 65% innerhalb der oberen 23 Ergebnisse. Innerhalb der ersten 17 Ergebnisse liegt die durchschnittliche Suchgenauigkeit jedoch nie unter 80%. Folglich schneidet TSISQ etwas schlechter ab als in der Pilotstudie (vgl. Koukal et al., 2013) (75%), bestätigt aber dennoch die insgesamt bemerkenswerten Ergebnisse.

Die Ergebnisse dieser Studien zeigen insgesamt, dass TSISQ sehr gut dazu geeignet ist, wissenschaftliche Literaturrechercheprozesse zu verbessern, insbesondere bei der Verwendung von Volltext in natürlicher Sprache mit inhaltlichem Bezug zur Zieldomäne als Input. In allen Suchszenarien generiert die TSISQ-Volltextsuche qualitativ mindestens gleichwertige, häufig auch höherwertige Ergebnisse hinsichtlich der Relevanz der gefundenen Artikel. TSISQ eignet sich zudem, wie zu Beginn gefordert, am besten für die Formulierung von Suchanfragen in natürlicher Sprache. Zum Beispiel liefert ein vollständiger Abstract als Suchanfrage bei AISel, unserem Benchmark für eine stichwortbasierte Suchmaschine, entweder keine Ergebnisse (bei der Anwendung von „AND“-Bedingungen), oder die Ergebnisse bestehen (bei der Anwendung von „OR“-Bedingungen) aus einer völlig zufälligen Auswahl an Dokumenten aus den verschiedensten Forschungsgebieten, während TSISQ (mit den o.g. Einschränkungen) sinnvolle Ergebnisse liefert.

Bei LSI, dem Verfahren auf dem TSISQ basiert, handelt es sich um eine Methode aus dem Gebiet des Information Retrieval zur Identifikation semantischer Ähnlichkeiten, es ist hauptsächlich für die Suche nach semantischen Konzepten gedacht. Daher kann die klare Empfehlung gegeben werden, statt Stichwörtern fachgebietsspezifische Volltexte in natürlicher Sprache für die Suchanfragen zu verwenden. Aus der Erkenntnis, dass die Ergebnisse bei z.B. Abstracts als Suchanfrage thematisch spezifischer sind, während sie bei Verwendung von vollständigen Forschungsarbeiten diffuser bzw. weniger fokussiert ausfallen, folgt für ein optimales Ergebnis die Empfehlung, einen iterativen, zyklischen Prozess bei der Literaturrecherche einzuhalten. Die erste Suchanfrage sollte aus einem vollständigen Artikel des gewünschten Themengebiets bestehen, um eine semantisch „unschärfere“ Auswahl ähnlicher Paper als Ergebnis zu erhalten. Anschließend folgt eine manuelle Überprüfung der Ergebnisse zur Identifikation der relevantesten Artikel für eine spezifischere thematische Eingrenzung. Der nächste Schritt umfasst das Erstellen einer Sammlung prägnanter Textbausteine, Sätze und/oder Abschnitte, die in der vorab identifizierten Literaturauswahl enthalten sind. Im letzten Schritt wird die daraus resultierende Sammlung als Suchanfrage verwendet. Der Prozess sollte zyklisch wiederholt werden, bis keine neue, relevante Literatur mehr gefunden wird.

In Bezug auf das übergeordnete Forschungsziel, mit TSISQ die Effizienz wissenschaftlicher Literaturrechercheprozesse mithilfe von Methoden des Data Mining und Big Data Analytics zu erhöhen, können folgende Schlüsse gezogen werden:

- Die Anwendung des hier dargestellten Ansatzes hilft dabei, wertvolle Zeit beim Auffinden der relevanten Literatur zu einem beliebigen Forschungsfeld einzusparen.
- Auch wenn genügend Zeit vorhanden ist kann TSISQ dennoch dazu beitragen, die Vollständigkeit und die Effizienz von Literaturrecherchen zu steigern: Zum einen durch die im Vergleich zu menschlichen Fähigkeiten deutlich höhere Arbeitsgeschwindigkeit und zum anderen durch die Identifikation von Quellen, die andernfalls möglicherweise nicht gefunden oder berücksichtigt worden wären.
- Die im Rahmen der Literaturrecherche eingesparte Zeit kann für eine gründlichere Durchführung anderer Forschungsaufgaben aufgewendet und damit die Qualität der Forschung insgesamt erhöht werden.

Abschließend kann festgehalten werden, dass der im Rahmen dieses Projektes entwickelte Ansatz zur Optimierung der wissenschaftlichen Literaturrecherche eine nützliche Ergänzung zu den etablierten Suchmaschinen und -methoden in der Wissenschaft darstellt. Er kann im Rahmen jeder umfassenderen, etablierten Forschungsmethodik zum Einsatz kommen und dabei deutliche Effizienzgewinne bei der Literaturrecherche erzielen.

Operative Entscheidungsunterstützung in der Automobilindustrie mit Künstlicher Intelligenz:

In zweiten Hauptteil der vorliegenden Dissertation (Kapitel 3) werden die Ergebnisse und Erkenntnisse aus insgesamt drei Forschungsarbeiten zusammengefasst, die dem oben genannten Themenkomplex der Entscheidungsunterstützung in der Automobilindustrie durch Data Mining, Big Data Analytics, Business Intelligence und KI bzw. KNN angehören. Diese Veröffentlichungen sind das theoretische Fundament eines 4-jährigen Kooperationsprojektes mit einem großen deutschen Automobilhersteller. Dabei ist eine Sammlung an (Software-) Tools zur Entscheidungsunterstützung im Restwert- und Risikomanagement entstanden. Auch diese Veröffentlichungen bauen aufeinander auf. Die folgenden Themen und Publikationen werden in Teil B der vorliegenden Doktorarbeit adressiert:

- Gleue, C., Eilers, D., von Mettenheim, H.-J. und Breitner, M. H. (2017). “Decision Support for the Automotive Industry: Forecasting Residual Values Using Artificial Neural Networks”, veröffentlicht in den Proceedings der Wirtschaftsinformatik-Jahrestagung (WI 2017).
- Gleue, C., Eilers, D., von Mettenheim, H.-J. und Breitner, M. H. (2018). “Decision Support for the Automotive Industry: Forecasting Residual Values Using Artificial Neural Networks”, Veröffentlichung voraussichtlich in der Business & Information Systems Engineering in 2019, Online-Vorveröffentlichung über Springer (2018).

- Eilers, D., Köpp, C., Gleue, C. und Breitner, M. H. (2017). “It’s Not a Bug, It’s a Feature: How Visual Model Evaluation Can Help to Incorporate Human Domain Knowledge in Data Science”, veröffentlicht in den Proceedings der International Conference on Information Systems (ICIS 2017).

Die Herausforderung eines großen deutschen Automobilherstellers bestand vor einigen Jahren darin, bereits bei Abschluss von Leasingverträgen eine möglichst genaue Vorstellung über den aktuellen Marktwert und den zukünftigen Restwert seiner Fahrzeuge zu entwickeln. Diese Ausgangslage motivierte uns, das im Folgenden dargestellte Projekt gemeinsam mit dem Hersteller durchzuführen, mit dem Ziel, Risiken, Rückstellungen, Leasingraten und -angebote genauer berechnen und dementsprechend besser steuern zu können.

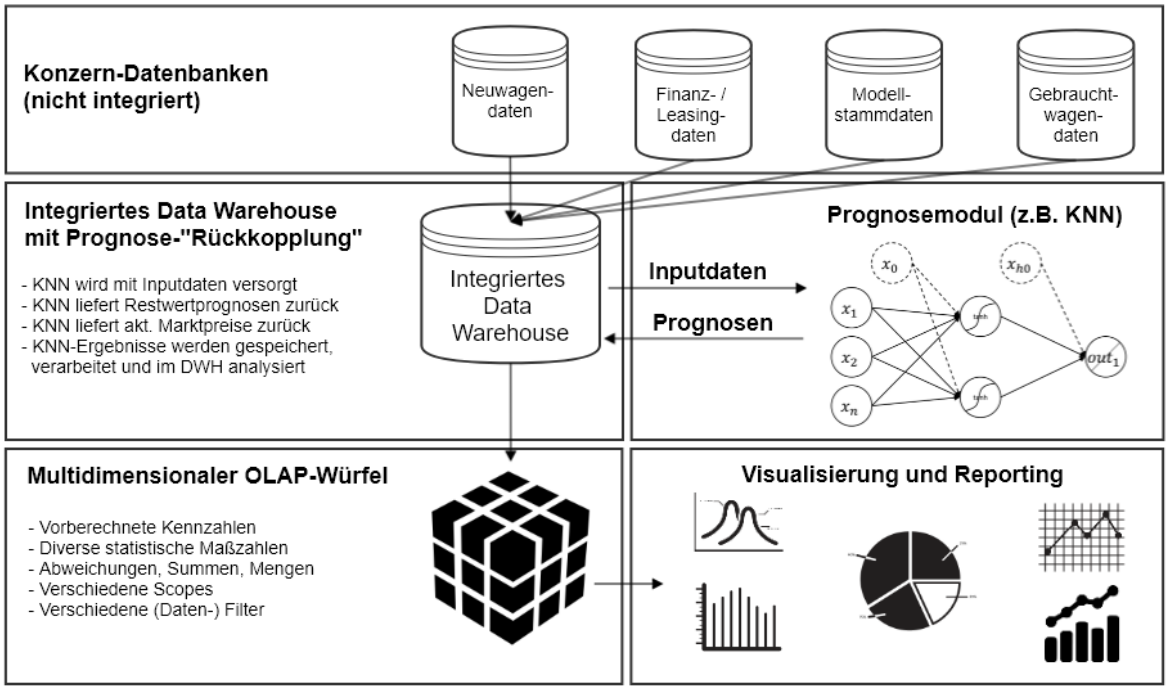

Abbildung 5: EUS für Markt- und Restwertprognosen im Leasinggeschäft

Bereits Jahre zuvor hatte der Konzern sich bereits darum bemüht, die dafür notwendige Datenbasis zu schaffen, indem allen Vertragshändlern starke finanzielle Anreize geboten wurden, alle ihre Gebrauchtfahrzeugverkäufe inkl. Fahrzeuginformationen zu erfassen und die entsprechenden Transaktionsdaten zu historisieren. Diese Daten galt es nun mithilfe der neuesten Erkenntnisse und Methoden aus den Gebieten der Entscheidungsunterstützung, Business Intelligence, Data Mining, Künstlicher Intelligenz und Big Data Analytics zielorientiert mit Daten aus zahlreichen weiteren Konzernsystemen zu integrieren, sie zu bereinigen, zu analysieren und entsprechend aufzubereiten. Nach einer ersten „Proof of Concept“-Phase und einer knapp 4-jährigen Projektlaufzeit unter Beteiligung diverser

Fachabteilungen, dem konzerneigenen BI-Kompetenzcenter und dem Institut für Wirtschaftsinformatik ist so ein intelligentes EUS entstanden, das in Abbildung 5 schematisch dargestellt ist.

Die Ergebnisse dieses Projekts lassen sich zusammenfassen wie folgt:

1. Wir stellen den exemplarischen Aufbau des im Projekt entstandenen EUS mit dem zuvor beschriebenen Zweck vor (vgl. Abbildung 5).
2. Der Einfluss der uns zur Verfügung stehenden Inputfaktoren auf die Markt- und Restwerte wurde untersucht und zahlreiche Erkenntnisse daraus abgeleitet.
3. Wir beschreiben unseren Ansatz zur Verwendung von KNN für eine Prognoseanwendung in diesem Anwendungsgebiet.

Darüber hinaus mussten verschiedenste, charakteristische Herausforderungen bewältigt werden: Der Umgang mit kurz- und insbesondere langfristigen Prognosehorizonten, die für genaue Ergebnisse notwendige, mehrstufige Methodik zur Bereinigung verrauschter Daten „aus der realen Welt“ sowie die Behandlung von zeit- und konjunkturabhängigen Variablen.

Eines der zunächst insbesondere von Fachexperten eher unerwarteten, empirischen Ergebnisse ist der Nachweis, dass potenzielle externe Einflussfaktoren wie Öl- bzw. Kraftstoffpreise, diverse makroökonomische Indizes wie DAX, der Ifo-Geschäftsklimaindex oder auch das BIP, der EZB-Leitzins oder die Arbeitslosenrate keinen signifikanten Einfluss auf die Markt- und Restwerte ausüben. Im Gegensatz dazu wurden die internen, vom Hersteller selbst beeinflussbaren Faktoren wie Fahrzeugalter und Kilometerleistung zum Zeitpunkt der Wiedervermarktung, Ausstattung des Fahrzeugs, Motorisierung oder Getriebeart als maßgebliche Einflussfaktoren identifiziert, was abermals den Wert eines solchen internen „Datenschatzes“ unterstreicht. Die letztlich aus unserer Forschung ableitbaren Erkenntnisse sind für den Hersteller besonders wertvoll, da auf deren Basis bereits frühzeitig, durch eine entsprechende Steuerung der Leasingangebote und damit des späteren Leasingportfolios[1], Fahrzeuge so konfiguriert werden können, dass sie sich zu einem späteren Zeitpunkt am Markt möglichst restwertstabil verhalten.

Weitere Erkenntnisse gab es auch bzgl. des Einflusses zeitbezogener Faktoren auf die Prognosegenauigkeit. Anhand des unseren Analysen zugrundeliegenden Datensatzes, der in dieser Form in der Forschung einzigartig ist, konnten wir nachweisen, dass die Marktpreise von Gebrauchtfahrzeugen sowohl saisonalen Effekten als auch einem über die Zeit sinkenden Trend

[1] Anm. des Verfassers: Nach Möglichkeit wäre sogar eine auf Restwertstabilität gesteuerte Produktion denkbar.

unterliegen, der, nachdem der preissenkende Effekt eines mit der Zeit stetig steigenden Modellalters bereits in der Modellbildung berücksichtigt wurde, auf die gesamtwirtschaftliche Situation im Betrachtungszeitraum zurückzuführen ist.
Bzgl. der Saisonalitäten konnten zum Ende eines Jahres generell niedrigere Marktwerte als zu Beginn beobachtet werden (vgl. Abbildung 6).

Es wurde zudem in einem Benchmark gezeigt, dass die letztlich auch operativ implementierte, auf KNN basierende Prognoseanwendung leistungsfähiger ist und deutlich genauere Ergebnisse liefert als lineare Modelle (Ridge Regression). Eine der großen Herausforderungen bei der technischen Umsetzung war dabei die Berücksichtigung des Faktors „Zeit“, bzw. der Variablen mit einem zeitlichen Bezug. Die letztlich von uns favorisierte Lösung besteht aus einer Kombination von KNN, um die nicht-linearen Beziehungen innerhalb des Basisdatensatzes abbilden zu können und einer linearen, abwärts gerichteten Adjustierung der Ergebnisse abhängig von dem gewünschten Prognosehorizont. Unsere Tests haben dabei gezeigt, dass diese lineare Anpassung der KNN-Ergebnisse (bis zu 1,5 Jahre in die Zukunft) keine nennenswerte Verzerrung verursacht und im Anschluss zumindest im Vergleich zu anderen von uns getesteten Methoden eine „Best Practice“ darstellt.

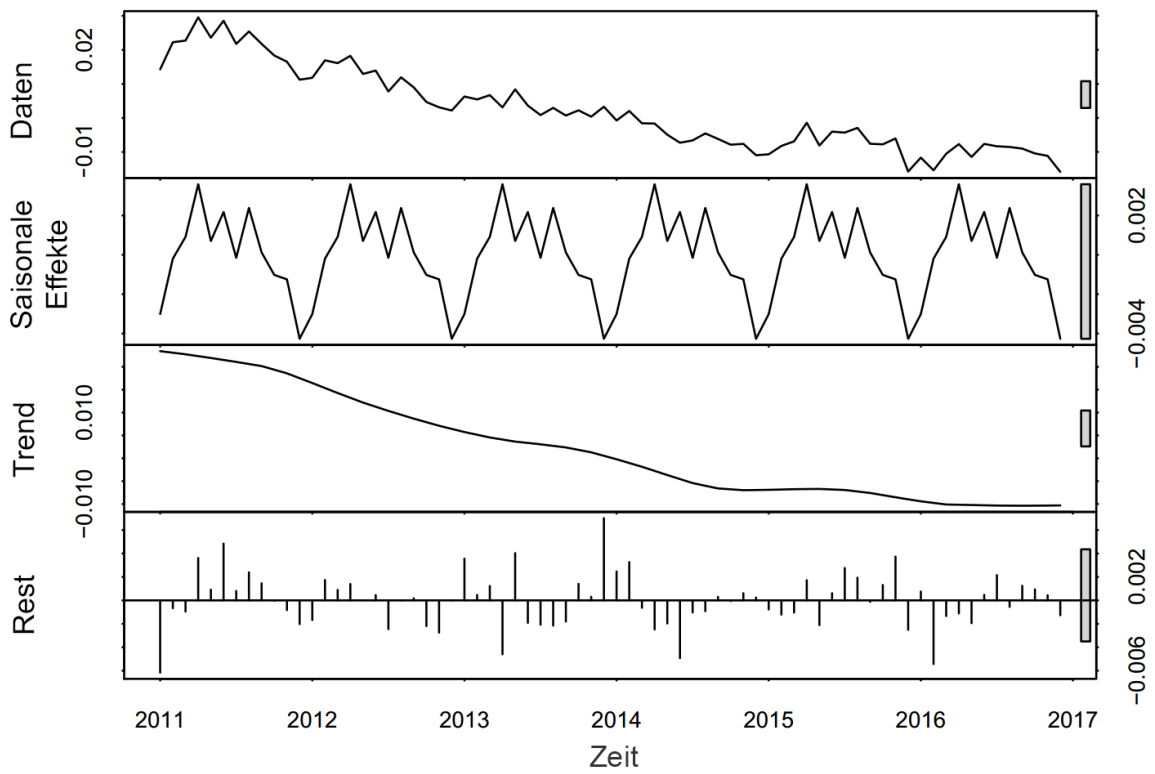

Abbildung 6: Zeitreihenzerlegung: Saisonale Effekte und Trends

Es wurde zudem untersucht, wie ein auf dieser Technologie und Methodik basierendes EUS in der wirtschaftlichen Praxis operationalisiert werden kann. Die gewonnenen Erkenntnisse und

Prozessoptimierungen wurden inzwischen bei dem Automobilhersteller mit Erfolg implementiert. Neben der Schaffung von tagesaktueller Transparenz am Gebrauchtwagenmarkt stand dabei die möglichst genaue Vorhersage der Zukunftsrestwerte im Vordergrund, wobei das Prognosemodul letztlich den entscheidenden Bestandteil des insgesamt deutlich umfangreicheren EUS ausmacht. Die mithilfe dieses EUS geschaffene Markttransparenz und die nachweislich deutlich verbesserte Prognosequalität sind ausschlaggebend für das Risikomanagement, denn eine systematische Verzerrung der kalkulierten Markt- und Restwerte in beiden Richtungen (über- oder unterschätzte Restwerte) bringt kurz- mittel- und langfristig negative Konsequenzen für Händler und Hersteller mit sich, beeinträchtigt die Wettbewerbsfähigkeit bzw. die Höhe der notwendigen Verkaufshilfen für den Vertrieb und/oder hat einen maßgeblichen Einfluss auf die Handelsmargen bei Wiederverkauf der Fahrzeuge. Somit ist eine bedarfsgerechte Rückstellungsbildung und marktgerechte Preisgestaltung im Leasinggeschäft von gleichermaßen hoher (wirtschaftlicher und strategischer) Bedeutung für Automobilhersteller und -händler.

Das Hauptziel dieses Projektes, die Etablierung eines rein datengetriebenen Ansatzes zur Restwertprognose von Leasingfahrzeugen, wurde erreicht und führte durch enge Zusammenarbeit zwischen dem Institut für Wirtschaftsinformatik, dem Fachbereich Finanz und dem Entwicklungsteam des Business Intelligence Competence Centers (BICC) des Automobilherstellers zu gesteigerter (Technologie-) Akzeptanz und Vertrauen innerhalb des Konzerns. Die Ergebnisse zeigen, dass KNN für die in der Praxis häufig verrauschten und unstrukturierten Daten gut geeignet sind. Dennoch ist eine gründliche Datenbereinigung und -aufbereitung unabdingbar, um gute und verlässliche Ergebnisse zu erzielen. Es wurde gezeigt, wie KNN zur Vorverarbeitung und Filterung der Datenbasis verwendet werden und wie ein Prognosemodell auf Basis eines KNN-Ansatzes als „Herzstück“ eines auf die Bedürfnisse der Finanzabteilung des Automobilherstellers maßgeschneiderten EUS entworfen werden kann.

III. Inhaltsverzeichnis

I. Abstract II

II. Management Summary IV

III. Inhaltsverzeichnis XVI

IV. Abbildungsverzeichnis XVIII

V. Tabellenverzeichnis XVIII

VI. Abkürzungsverzeichnis XIX

VII Überblick über Publikationen und Aufgabenverteilung 1

1. Einleitung 6

1.1 Problemstellung, Relevanz und Motivation 6

1.2 Aufbau der Arbeit und wissenschaftlicher Beitrag 7

2 Semantische Suche im Literaturrechercheprozess 9

2.1 Motivation und Zielsetzung 10

2.2 Literaturrecherche und Methoden des Information Retrieval 13

2.2.1 Stand der Forschung 13

2.2.2 Forschungsmethodik 19

2.3 Optimierung des Literaturrechercheprozesses mit TSISQ 21

2.3.1 Theoretischer Hintergrund und Funktionsweise 22

2.3.2 Implementierung, Architektur und Benutzeroberfläche 25

2.4 Demonstration, Evaluation und Benchmark von TSISQ 30

2.4.1 Benchmark mit einer stichwortbasierten Suchmaschine 30

2.4.2 Identifikation der falsch-positiven und falsch-negativen Fehlerrate 33

2.5 Kritische Würdigung und Diskussion der Ergebnisse 35

2.6 Fazit 41

3 Operative Entscheidungsunterstützung mit KI 44

3.1 Gesamtwirtschaftlicher Kontext, Motivation und Projektziele 45

3.1.1 Wirtschaftliche Bedeutung des Leasinggeschäfts für Automobilhersteller 46

3.1.2 Markt- und Restwertprognosen in der wissenschaftlichen Literatur 47

3.1.3 Projektziele 49

3.2 Gesamtsystem zur Entscheidungsunterstützung im Leasinggeschäft 51

3.2.1 Systemarchitektur und Datenquellen 52

3.2.2 Inputfaktoren für die Markt- und Restwertprognose 56

3.2.3 Markttrends und saisonale Effekte 57

3.2.4 Der Einfluss externer Faktoren auf Fahrzeugrestwerte 59

3.3 Das Prognosemodul als Kern des EUS 61

3.3.1 KNN für die Restwertprognose: Eignung und Funktionsweise 63

3.3.2 Erkennung von Ausreißern und Datenbereinigung 65

3.3.3 Vorhersagekraft der Inputfaktoren und Prognosequalität 66

3.3.4 Prognoseverfahren: Anwendungsszenarien und Benchmark 68

3.4 Kritische Würdigung und Diskussion der Ergebnisse 72

3.5 Fazit 76

4 Kritische Reflexion und Ausblick 78

5 Fazit 85

6 Literaturverzeichnis 86

7 Appendix ..... A

A: Enhancing Literature Research Processes: A Glance at an Approach Based on Latent Semantic Indexing ..... A

B: Enhancing literature Review Methods - towards More Efficient literature Research with Latent Semantic Indexing ..... B

C: Enhancing Literature Review Methods - Evaluation of a Literature Search Approach based on Latent Semantic Indexing C

D: Decision Support for the Automotive Industry: Forecasting Residual Values Using Artificial Neural Networks (WI) ..... D

E: It’s Not a Bug, It’s a Feature: How Visual Model Evaluation Can Help to Incorporate Human Domain Knowledge in Data Science ..... E

F: Decision Support for the Automotive Industry: Forecasting Residual Values Using Artificial Neural Networks (BISE) ..... F