Diplomarbeit

A Procedural Model for Extracting Knowledge from Industrial Maintenance Reports

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines

Diplom-Ingenieurs

unter der Leitung von

Univ.-Prof. Dr.-Ing. Dipl. Wirt.-Ing. Prof. eh. Dr. h.c. Wilfried Sihn
(E330 Institut für Managementwissenschaften, Bereich: Betriebstechnik und Systemplanung)

Dr.-Ing. Fazel Ansari, MSc
(E330 Institut für Managementwissenschaften, Bereich: Betriebstechnik und Systemplanung,

Forschungsgrupe Smart and Knowledge-Based Maintenance)

eingereicht an der Technischen Universität Wien

Fakultät für Maschinenwesen und Betriebswissenschaften

von

Johannes Breitschopf

01226629 (066 482)

Harrstraße 10

4407 Dietach

Wien, im Februar 2020

Johannes Breitschopf

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.

TECHNISCHE UNIVERSITÄT WIEN

Ich habe zur Kenntnis genommen, dass ich zur Drucklegung meiner Arbeit unter der Bezeichnung

Diplomarbeit

nur mit Bewilligung der Prüfungskommission berechtigt bin.

Ich erkläre weiters Eides statt, dass ich meine Diplomarbeit nach den anerkannten Grundsätzen für wissenschaftliche Abhandlungen selbstständig ausgeführt habe und alle verwendeten Hilfsmittel, insbesondere die zugrunde gelegte Literatur, genannt habe.

Weiters erkläre ich, dass ich dieses Diplomarbeitsthema bisher weder im In- noch Ausland (einer Beurteilerin/einem Beurteiler zur Begutachtung) in irgendeiner Form als Prüfungsarbeit vorgelegt habe und dass diese Arbeit mit der vom Begutachter beurteilten Arbeit übereinstimmt.

Wien, im Februar 2020

Johannes Breitschopf

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Danksagung

An dieser Stelle möchte ich mich bei all denjenigen bedanken, die mich bei der Erstellung dieser Arbeit unterstützt haben. Zuerst gilt mein Dank dem Betreuer dieser Bachelorarbeit, Herrn Dr. Fazel Ansari, der durch Anregungen und konstruktive Kritik bei der Erstellung maßgeblich mitgewirkt hat. Weiters möchte ich mich bei Alexander Aigner für die Korrekturen in der sprachlichen Umsetzung bedanken. Zuletzt gilt mein Dank meinen Eltern Johann und Ludmilla Breitschopf, durch deren Unterstützung mein Studium überhaupt erst ermöglicht wurde.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Abstract

With the emergence of IoT platforms, sensing technologies and condition monitoring systems, data-driven approaches have flourished driving a new era in industrial maintenance. Although the developed knowledge-based models have achieved remarkable advances in maintenance scheduling and decision support, they are mainly limited to the analysis of recorded mechanical conditions and equipment-related data.

Conversely, this work focuses on the rather neglected side of human background knowledge, which is expressed and recorded by uncodified textual data. It strives for enhancing existing data-driven approaches by codifying and explicating latent human knowledge with appropriate text-mining methods.

The investigated concepts of association measuring and sentiment analysis are embedded in a comprehensive procedural model for knowledge extraction involving necessary steps for data preparation as well as the visualization of results. The proposed approach is finally applied to a use-case in the industrial sector for analyzing textual maintenance reports. The obtained outcome reveals inferences such as failure/action pairs and expresses maintenance staff's opinions towards occurring events.

As part of future works the presented approach could serve as foundation for deriving text- and knowledge-based Key Performance Indicators (KPI) in the environment of industrial maintenance. The extracted opinions of experts towards equipment as well as the strength of association with respect to occurring failures and events could be utilized to complement established performance indicators, such as Overall Equipment Effectiveness (OEE). Moreover, the approach could be integrated in the reporting function of "Digital Shift Books", which ensure standards in textual recording of maintenance events and thus build a valuable textual data source for knowledge mining.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Kurzfassung

Die derzeitige Entwicklung von IoT-Plattformen, Sensortechnologien und Condition-Monitoring Systemen ermöglicht datengetriebene Ansätze, die eine neue Sichtweise in der industriellen Instandhaltung bewirken. Obwohl die wissensbasierten Modelle der Instandhaltung bereits bemerkenswerte Fortschritte bei der Planung und der Entscheidungsunterstützung erzielt haben, beschränken sie sich hauptsächlich auf die Analyse aufgezeichneter Zustände und maschinenbezogener Daten.

Im Gegensatz dazu konzentriert sich diese Arbeit auf die eher vernachlässigte Seite des menschlichen Hintergrundwissens, welches häufig über unstrukturierten Textdaten kommuniziert und festgehalten wird. Ziel ist es, bestehende datengetriebene Ansätze zu verbessern, indem latentes menschliches Wissen mit geeigneten Text-Mining-Methoden zugänglich gemacht und expliziert wird. Die untersuchten Konzepte Assoziationsanalyse und Stimmungsanalyse sind in einem umfassenden Vorgehensmodell zur Wissensgewinnung eingebettet, das notwendige Schritte zur Datenaufbereitung sowie die Visualisierung der Ergebnisse beinhaltet. Das vorgeschlagene Konzept wird schließlich auf einen Anwendungsfall im industriellen Bereich zur Analyse von Wartungsberichten angewendet. Die erhaltenen Ergebnisse lassen Rückschlüsse auf Fehler und Wartungstätigkeiten zu und geben die Meinung des Wartungspersonals zu extrahierten Entitäten wieder.

Im Rahmen zukünftiger Arbeiten kann der vorgestellte Ansatz als Grundlage für die Ableitung text- und wissensbasierter Key Performance Indicators (KPI) im Umfeld der industriellen Instandhaltung dienen. Die extrahierten Meinungen von Experten zu Anlagen, sowie die Ausprägung der Assoziation in Bezug zu auftretenden Fehlern und Ereignissen könnten genutzt werden, um etablierte Leistungsindikatoren wie die Gesamtanlageneffektivität (OEE) zu ergänzen. Darüber hinaus könnte der Ansatz in die Berichtsfunktion von „Digital Shift Books“ integriert werden, die Standards für die Textaufzeichnung von Wartungsereignissen sicherstellen und so eine wertvolle Textdatenquelle für das Extrahieren von Wissen bilden.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Table of contents

1 Introduction 1
1.1 Industrial Maintenance and Text-Mining 1
1.2 Problem Statement, Objective and Research Question 4
1.3 Thesis Structure 6
2 Theoretical Background 8
2.1 Industrial Maintenance 8
2.1.1 Objectives of Industrial Maintenance 9
2.1.2 Maintenance Strategies 10
2.1.3 Knowledge-based Maintenance 16
2.2 Text-Mining 20
2.2.1 Text Pre-processing 23
2.2.2 Word Embedding and Representation 37
2.2.3 Association Measuring 43
2.2.4 Opinion-Mining and Sentiment Detection 46
3 Practical Application of Text Mining: State-of-the-Art 53
3.1 Text-Mining Approaches applied in Various Business Sectors: A Brief Overview 53
3.2 Applications of Text-Mining in Production Systems 56
3.3 Text-Mining Approaches in Industrial Maintenance 59
3.4 Summary of the literature review 66
4 Applied Research Design Methodology 68
4.1 Model Development with Design Science Methodology 68
4.2 Statistical Evaluation 71
4.2.1 Classification Evaluation 71
4.2.2 Sample Selection and Sample Size Definition 73
5 Design and Implementation of Procedural Model for Knowledge Extraction from Textual Maintenance Records 75
5.1 Use-Case Description 75
5.2 Overview of Procedural Model for Knowledge Extraction 76
5.2.1 Pre-processing, Word embedding and Keyword extraction 78
5.2.2 Implementation of Association Measuring 85

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

5.2.3 Implementation of Opinion-mining and Sentiment Detection 88
5.3 Model Evaluation 91
6 Summary and Consolidation of Key Findings.....94
6.1 Results of Applied Methods 94
6.2 Results with Respect to Research Questions 99
7 Outlook and Limitations.....104
8 References.....109
9 List of Figures.....118
10 List of Formulas.....120
11 List of Tables.....121
12 List of Abbreviations.....122

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

1 Introduction

1.1 Industrial Maintenance and Text-Mining

Industrial maintenance has become a crucial part in the cost structure of goods and make up between 15 and 60 percent of the operating costs, depending on asset intensity and industrial sector.[1] It has been detected as an important factor for the direct increase of plant profitability and is no longer regarded as a "necessary evil".[2] Apart from the image and reputation of industrial maintenance changing from a main cost-driver to an essential enabler of efficiency enhancement, the scope of duties has been reconsidered extending the traditional tasks of service, inspection and repair by the duty of plant and safety improvement.[3] Nowadays, the responsibility of maintenance operators goes beyond the mere arrangement of repair and spare part procurement; maintenance staff is expected to act anticipatorily and proactively in order to avoid machinery downtime and to ensure safety.[4]

According to surveys of maintenance management effectiveness, 33 percent of all maintenance costs can be traced back to premature change of machinery components or delayed maintenance actions resulting in resource inefficiency and unexpected downtime.[5] The development of sensing and computational technologies, the access to huge amount of data, as well as data-driven processes have reinforced the rethinking of maintenance approaches and have pushed the evolution of knowledge-based systems attempting to offer advanced solutions for the aforementioned difficulties.[6] Especially the application of real-time based technologies connecting physical devices via Internet, which is also referred to as Internet of Things (IoT), allows new approaches in the area of maintenance and is deemed to have high potential for satisfying individual customer requirements in highly competitive markets.[7]

In 2018 the global market of Predictive Maintenance solutions reached a volume of 3.3 billion dollars and is expected to expand with a compound annual growth rate of 39%, as shown in Figure 1. The number of Predictive Maintenance vendors has doubled within the last two years saving organizations worldwide about 17 billion dollars in 2018. Companies thus report great ROI for PdM projects coming along with typically 10%-50% reduction in maintenance costs.[8]

[1] cf. Mobley, 2002, p.1
[2] cf. Ibid., p.1
[3] cf. Ansari/Glawar, 2019, p.318
[4] cf. Reichel/Müller/Haeffs, 2018, V
[5] cf. Mobley, 2002, p.1
[6] cf. Ansari/Glawar/Nemeth, 2019, p.1
[7] cf. Biedermann, 2018, p.5
[8] cf. IoT Analytics, 2019, p.2

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

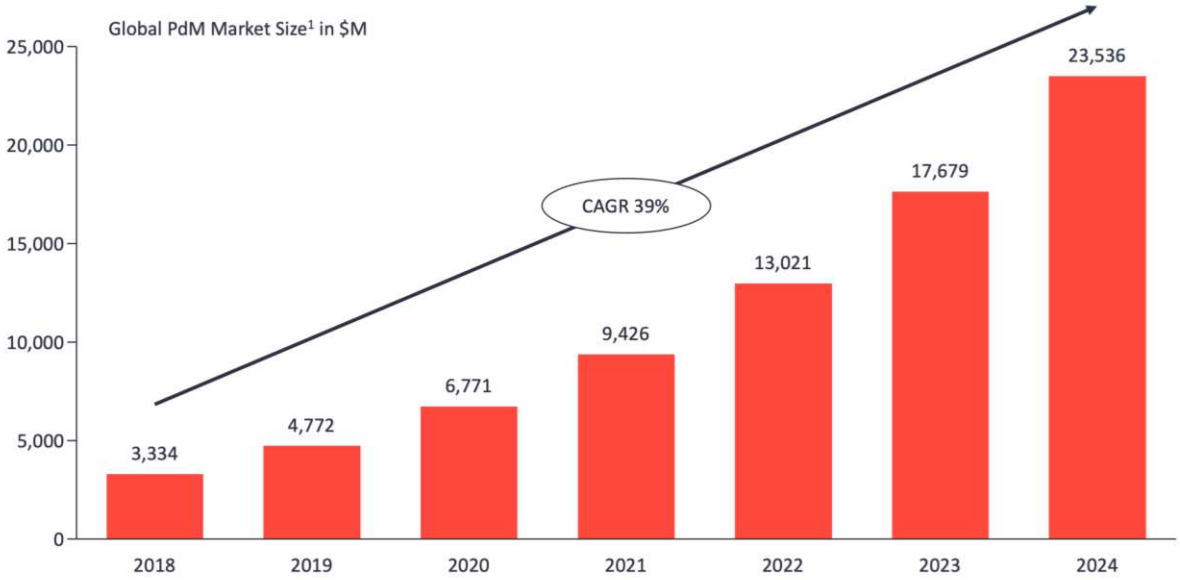

Figure 1: Global Market Development Predictive Maintenance[9]

The main goal of Predictive Maintenance includes the prediction of failures in order to reduce downtime, process instability and resource instability striving for a shift from fixed scheduled to operating-condition-based events. However, analyzing troubles in maintenance reveals mismanagement, such as inappropriate selection of strategies, as a root cause of failures, which in particular goes along with deficits in comprehensive and up-to-date knowledge. Overcoming these difficulties requires a knowledge-based decision-support and recommendation system, which extends the approach of failure prediction with the scheduling of specific events and their procedure of execution.[10] That so called prescriptive maintenance approach unites the fields of prediction, planning and decision-making and is expected to become the leading industrial application in the upcoming years.[11]

The key element of prescriptive maintenance and knowledge-based maintenance in general is the provided knowledge acquired from various data sources. The definition of KBM as a general concept of holistic consideration has been extended and categorized by introducing approaches in relation to different complexity and maturity levels as illustrated in Figure 9.[12] Nevertheless, apart from the aforementioned literature, data-driven maintenance concepts still focus on the monitoring of mechanical conditions and on equipment-related data which are collected by sensors and measuring instruments. For instance R. Keith Mobley suggests vibration-, thermography- and tribology monitoring as most promising testing techniques for generating a substantial data base.[13] Moreover, Moldaschl et al. focus on a wireless

[9] IoT Analytics, 2019, p.2
[10] cf. Ansari/Glawar/Nemeth, 2019, p.1
[11] cf. Lueth/Patsioura/Williams/Kermani, 2016, p.8
[12] cf. Ansari/Glawar/Nemeth, 2019, p.4
[13] cf. Mobley, 2002, p.99

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

RFID-based Sensor Network for recording states of machines in order to avoid unexpected downtime.[14] J. Shin and H. Jun propose Micro-Electro-Mechanical Systems and product-embedded devices to be appropriate technologies for condition monitoring and assessment of deterioration status.[15]

Although condition monitoring is an important and effective component of knowledge-based maintenance and sensor-recorded data a valuable source, corporate information is available in many different formats, which also includes unstructured data with irregularities and ambiguities. In industrial environments, approximately 80% of corporate information is obtainable in form of text documents, such as repair information, service reports, manufacturing quality documentation and customer help desk notes.[16] Apart from their common existence and availability, human created text documents are a very valuable and informative source. In contrast to sensor-recorded data, text implicitly contains human knowledge acquired from experiences, which could be extracted by suitable text-mining approaches.[17] Moreover, textual analysis in industry enables the disclosure of hidden relations and patterns which might remain undiscovered by solely regarding non-textual sources.[18] On the other side, textual data also comes along with difficulties due to its lack of regularity. The form and quality of manually drafted text documents depends on the individual authors and may contain irregularities such as typing errors which have to be corrected and cleansed for ensuring appropriate data quality. Mining this kind of data thus requires a sophisticated approach based on multiple pre-processing stages in order to prepare the data for further analysis.[19]

The discipline of mining texts also referred to as Knowledge-Discovery from Text can be classified as a sub-domain of data mining being characterized by the analysis of large databases for disclosing causes and patterns.[20] In contrast to data mining text mining processes unstructured and semi-structured data, such as e-mails and reports, aiming at the extraction of explicit but also implicit concepts.[21] The computational analysis and comprehension of unstructured natural language is a challenging task trying to imitate the human capabilities of language understanding, contextual perception and pattern recognition.[22] The tasks and challenges arise across all different layers of human language including semantic analysis for extracting the meaning of words and sentences, morphological processing for coping with word inflections and corresponding pre and suffixes as well as syntax tree

[14] cf. Moldaschl/Travnik/Frank/Bardong, 2019, p.255
[15] cf. Shin/Jun, 2015, pp.119f
[16] cf. Ur-Rahman, 2017, p.430
[17] cf. Ansari/Uhr/Fathi, 2014, p.15
[18] cf. Gupta/Lehal, 2009, p.60
[19] cf. Anandarajan/Hill/Nolan, 2019, p.14
[20] cf. Gupta/Lehal, 2009, p.60
[21] cf. Dörre/Gerstl/Seiffert/Fayyad, 1999, pp.398f
[22] cf. Gupta/Lehal, 2009, p.60

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

parsing for analyzing the sentence structure.[23] In order to cope with these challenges text-mining draws on methods from machine learning, information extraction and knowledge management being implemented in models of multiple layers.[24]

1.2 Problem Statement, Objective and Research Question

This master thesis focusses on the integration of text-mining into knowledge-based maintenance by extracting implicit knowledge as well as codifying uncodified explicit knowledge from industrial text documents and providing the output to associated domain experts. The investigated methods, association measuring and sentiment analysis, imitate the human ability of learning from text by classifying contents as well as building and weighting associations.

For the purpose of illustration, the methods are applied to a use-case in the industrial sector. Company A stores maintenance records which are categorized to predefined meta-data such as machinery components and provide information about the executed maintenance action. Notably, the name of the company is not mentioned due to data privacy reasons and is referred to as company A. The provided free-text reports are semi-structured text documents written by maintenance operators containing abbreviations, spelling mistakes, ambiguities as well as linguistic mixtures. In order to prepare the data for further analysis and avoid biased results, the proposed approach thus includes an extensive pre-processing section.

From a more general perspective, the superior objective is the exploitation of human-related knowledge sources extending existing concepts of condition monitoring and machinery data-driven maintenance. In particular, this work aims at creating a procedural model for overcoming the challenges of text processing and for accomplishing knowledge extraction with the concepts of association measuring and sentiment analysis. The expected outcome of the conducted analysis is the reveal of hidden information about equipment components and associated failures visualized and provided to maintenance experts. Hence, this work concentrates on mining textual data and should be seen as a contribution to the development of comprehensive knowledge-based maintenance approaches such as prescriptive maintenance strategies.

Following, the problem areas, objectives and research questions are listed, which arise in the field of industrial maintenance with a focus on knowledge discovery from text. The introduced issues are not only limited to the specific use-case but are dedicated to the intersection of text-mining and industrial maintenance in general.

[23] cf. Denham/Lobeck, 2010, p.139
[24] cf. Gupta/Lehal, 2009, p.60

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Problem areas

P1: Unstructured text documents are created but not exploited - Included domain knowledge as well as knowledge about influencing factors that lead to malfunctions is not extracted.

P2: The huge amount of textual data is difficult to be managed by humans and not suitable for manual analysis.

P3: Text documents contain ambiguous and incomplete information and are not directly accessible for computational analysis.

P4: Revealed information from textual maintenance data and derived recommendations are not coordinated with aspects of production management.

Objectives

O1: Building a qualitative procedural text-mining model as foundation for the development of corresponding software prototype.

O2: Selection of appropriate algorithms and their adaptions for extracting of latent patterns as well as explication of implicit knowledge from maintenance textual reports.

O3: Transparent and clear communication of the extracted results by visualization.

O4: Creating a recommendation system for standardizing text reports and improving quality of textual data written by machine operators.

Research questions:

Q1: Does the integration of text-mining significantly contribute to improvement and enrichment of prescriptive maintenance models?

Q2: Which methods succeed in explicating implicit knowledge and detecting hidden patterns in industrial maintenance?

Q3: How should the results be effectively communicated and visualized?

Q4: How to recommend standard reporting structures and templates including content for typical problems to improve quality of text reports?

Because of the wide range and complexity of associated problems in the field of industrial maintenance and text-mining, the thesis at first focuses on problem statements P1-P3, objectives O1-O3 and research questions Q1-Q3 which are elaborated in quantitative and qualitative approaches. Additionally, problem P4, objectives O4 as well as research questions Q4 are dealt with in qualitative

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

approaches for achieving a complete and comprehensible procedural model but remain open for more detailed investigation. The connections between the representatives of each category are illustrated in the figure below.

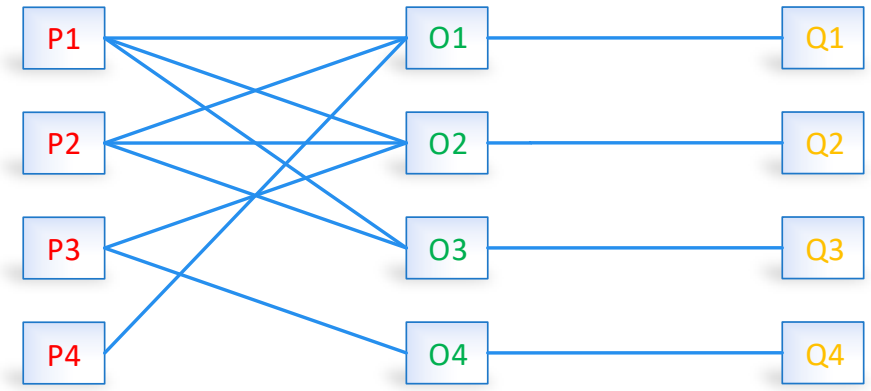

Figure 2: Relations between problem areas, objectives and research questions

1.3 Thesis Structure

The thesis is structured in six main parts: At first the theoretical background of industrial maintenance and text-mining is covered introducing a set of concepts and methods as general framework for the final use-case implementation. The section state of the art examines text-mining in different fields of application considering text-mining solutions and approaches of various problem areas. The general methodology for model creation and model evaluation is dealt with in section 4 providing tools and guidelines for sophisticated problem solving. In the subsequent chapter the use-case and the proposed procedural model are introduced, and the implemented methods exemplified by a step by step explanation. The visualized results are then discussed with respect to the use-case as well as from the perspective of the stated research questions. Finally, section 7 is dedicated to the limitations of the applied procedural model and provides recommendations for further enhancement. Figure 3 illustrates the structure of work and the logic connections between its chapters.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

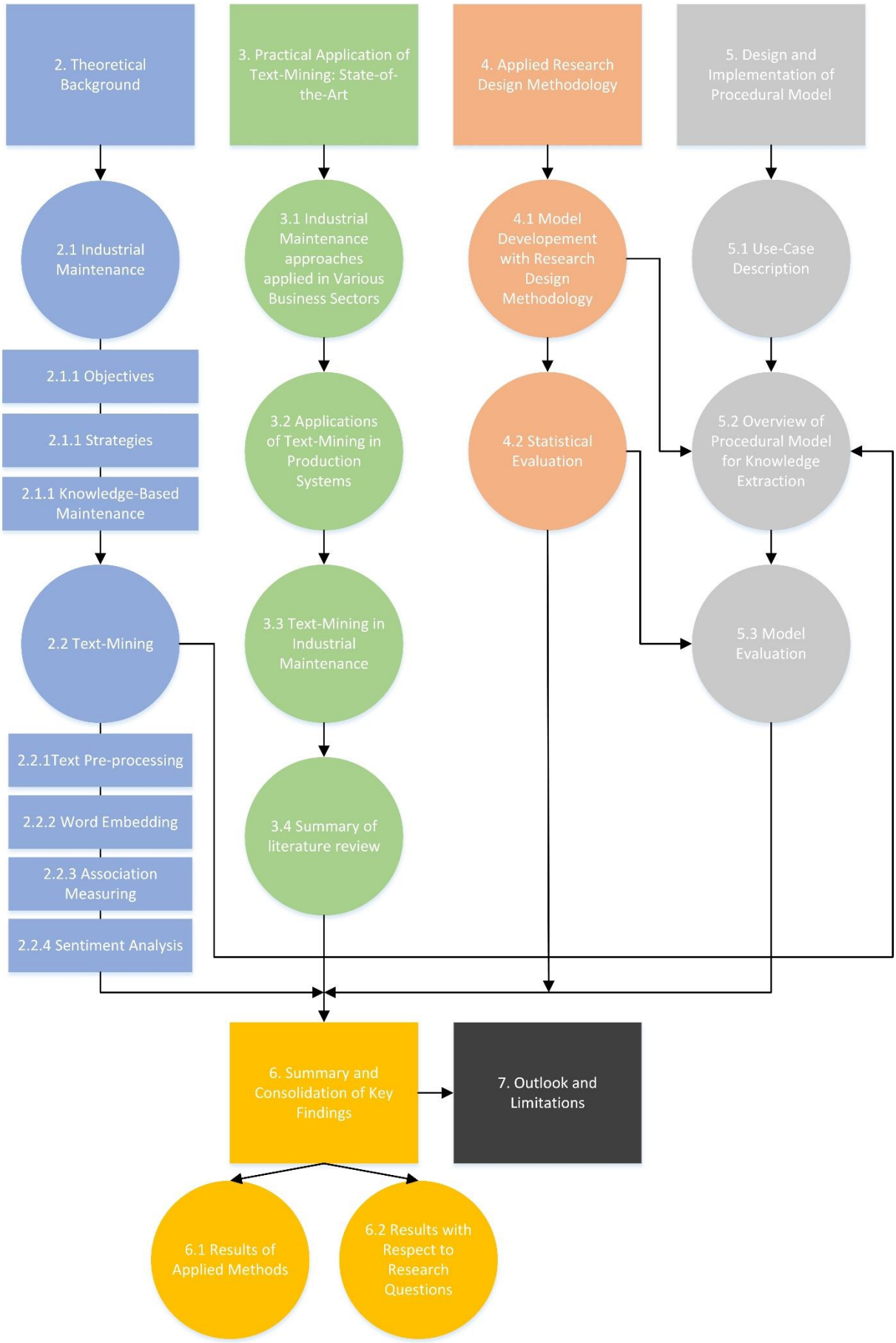

Figure 3: Depiction of thesis structure

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

2 Theoretical Background

2.1 Industrial Maintenance

Industrial maintenance is a main driver of a reliable and cost-efficient production and is crucial for guaranteeing process safety and equipment availability.[25] However, due to the increasing complexity of production and transport facilities including automated and chained processes, the scheduling and managing of maintenance actions have become more challenging. In order to avoid expensive downtime in production and risks concerning security and environment, maintenance activities are not limited to the repair of machinery break down anymore and require a data-driven and proactive strategy.[26]

According to DIN 31051 industrial maintenance is defined as the combination of all technical and administrative activities as well as management measures during the life cycle of an object, which preserve or restore the functional condition in order to satisfy all functional requirements.[27] In this context the term object may refer to entire plants and equipment but also to individual components.[28] As shown in Figure 4, DIN 31051 divides the tasks of maintenance into four basic preventive and corrective measures:

- Service: Preservation of the desired condition and reduction of wear e.g. lubrication
- Inspection: Observation of the actual state and analysis of wear causes and their corresponding consequences e.g. measuring the thickness of a brake disk
- Repair: Restoring the initial condition and the functional capability without improvement e.g. Replacement of a worn component by an equivalent part
- Improvement: Measures for improving functional safety without changing the function itself including technical, administrative and organizational actions. e.g. Disposing of weaknesses by constructive changes of components[29]

[25] cf. Reichel/Müller/Haeffs, 2018, p.V
[26] cf. Matyas, 2019, p.27
[27] cf. Deutsches Institut für Normung, DIN/31051, p.4
[28] cf. Schenk, 2010, p.23
[29] cf. Deutsches Institut für Normung, DIN/31051, pp.5f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

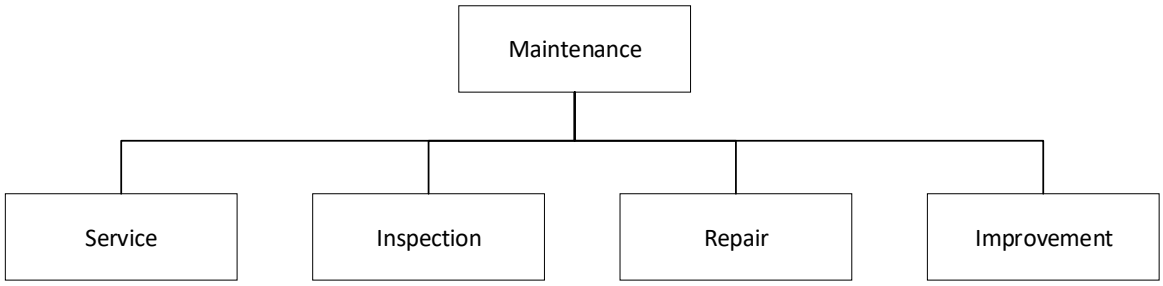

Figure 4: Basic measures of maintenance[30]

2.1.1 Objectives of Industrial Maintenance

In general, the superior goal of industrial maintenance and its basic measures is the minimization of the company's overall costs under consideration and improvement of safety as well as environmental impact.[31] It strives for an optimum, which balances the expenses of maintenance against the costs of machinery downtime including difficultly quantifiable influences such as delivery adherence of replacement parts.[32] In other words, a reduction in maintenance activities has benefits with respect to staff-, administrative- and material expenses, but leads to a rise in risk and down-time causing repair costs and efforts for compensating production irregularities. Furthermore, enhanced maintenance intensity has a positive impact on equipment availability striving for amortization of increased maintenance expenses.
Figure 5 illustrates the mentioned balancing act reaching its maintenance optimum at minimum total costs.

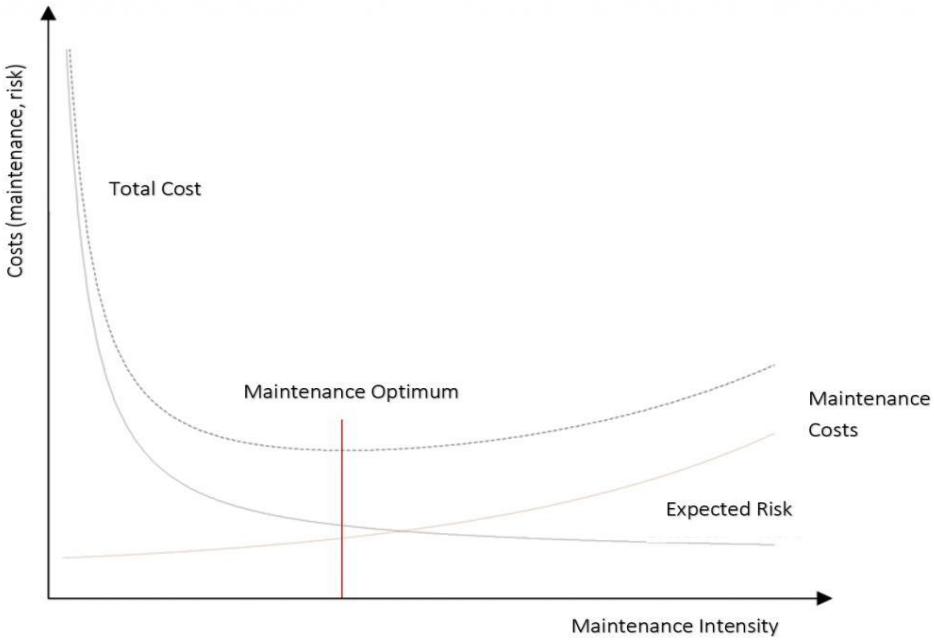

Figure 5: Maintenance optimum [33]

[30] vgl. Deutsches Institut für Normung, DIN/31051, p.4
[31] cf. Schenk, 2010, p.VII
[32] cf. Matyas, 2019, p.32
[33] URL: http://meliorate.de, 20.05.2019

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Companies' cost accounting departments are usually able to determine the total costs by aggregating expenses of planned and unplanned maintenance measures as well as costs caused by machine failure. Nevertheless, the current total costs correspond only to a single point in the graph of the function. Unless there is any further historical data available for interpolation, the run and parameters of the total cost curve cannot be determined. As a consequence, there is no detailed knowledge about the actual state and whether the current maintenance intensity should be increased, decreased or is situated at an optimal level. In order to overcome this difficulty, maintenance strategies and selection methods have been developed which are introduced in chapter 2.1.2.[34]

Apart from machinery availability and plant profitability, industrial maintenance aims at avoiding accidents and preventing environmental harm or even personal injury.[35] This includes law observation, maintenance documentation but also technical and constructive changes for safety improvement.[36] Because of that wide range of tasks and duties, industrial maintenance has to be seen as an integrated part of production and ought to be coordinated with logistical and manufacturing processes. Especially, the shift from functional-oriented to process-oriented approaches distinguishing between main-, management and assistance processes with a focus on customer orientation, blurred the lines of responsibility between production, maintenance, logistics and quality-management.[37] Considering for example chained lean production systems with minimized lead time, equipment availability and therefore industrial maintenance has become an important player being involved to typical tasks of logistics such as delivery reliability. In summary, industrial maintenance as integrated part affects the success across various production departments and thus pursues the superior objective of plant profitability and customer satisfaction.

2.1.2 Maintenance Strategies

In order to achieve defined objectives of industrial maintenance, it is necessary to pursue a strategy. A strategy is a systematic and structured procedure to accomplish goals, which may not be confused with the goals themselves.[38]
In the application field of industrial maintenance, strategies specify the measures, the objects and their time scheduling of conduction considering safety-, technical-, economical- as well as legislative restrictions.[39] With the development of sensors and monitoring systems, a new class of maintenance strategies emerged offering a data-

[34] cf. Matyas, 2019, p.49
[35] cf. Ibid., pp.40f
[36] cf. Schenk, 2010, p.VII
[37] cf. Matyas, 2019, p.27
[38] cf. Gudehus, 2012, p.107
[39] cf. Schenk, 2010, p.26

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

driven alternative to conventional rule based approaches.[40] Thus, taxonomy is separated in two categories; approaches with sensors and approaches without sensing technologies as follows:[41]

a.) Maintenance Approaches without Sensing Technologies

The **reactive maintenance** is one of the simplest approaches and is also called Run-to-failure strategy. Maintenance actions are conducted whenever failures occur without any kinds of measures in advance.[42] Planning and managing efforts are almost avoided, which is why it is still applied especially in small enterprises. Nevertheless, it has shown to be little efficient, since the costs of downtime and maintenance often exceed the expenses of preventive measures.[43]

The **preventive strategy** focusses on avoiding failures and risks concerning equipment, persons and environment. Objects and components are maintained in regular intervals, depending on fixed time periods or machining cycles.[44] Preventive strategies are advantageous for objects, which show steady failure behavior and of which lifetime can be predicted rather easily. Nevertheless, preventive maintenance measures are always conducted independently of the current equipment condition. Thus, in order to minimize cost and increase utilization the intervals are tried to be adapted to the lifetime of objects. It's obvious that a component change immediately prior failure is desirable with respect to utilization and down time reduction. Therefore, the discovery of the optimal interval has become a crucial task in preventive maintenance. It is mainly based on statistical data, despite facing the following challenges:

- Different average intervals between failures
- Different distribution of operating life
- Insufficient failure documentation
- Insufficient statistical failure experience[45]

In contrast to reactive and preventive methods, **proactive maintenance** concentrates on detecting and correcting root causes of failures.[46] It aims at avoiding and amending circumstances which lead to deterioration. In this approach analysis and investigation are not pointed at evaluating the degree of material degradation for realizing beginning states of failures; it targets disclosure and correction of abnormal

[40] cf. Matyas, 2019, p.300
[41] cf. Ansari/Glawar, 2019, pp.320f
[42] cf. Ibid., pp.320f
[43] cf. Susto/Schirru/Pampuri/McLoone/Beghi, 2015, p.812
[44] cf. Ansari/Glawar, 2019, pp.320f
[45] Matyas, 2002, p.14
[46] cf. Ansari/Glawar, 2019, pp.320f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

root causes being responsible for volatile operating conditions. Therefore, the triggers of failure are eliminated or averted.[47] One Example of proactive maintenance measures are training and instruction programs for avoiding inappropriate equipment handling and damage.

b.) Maintenance Approaches with Sensing Technologies

Condition-based maintenance relies on the current actual state of objects and components monitored by human documentation or sensing technologies.[48] It strives for the timely discovery of potential damages and represents the next level of maintenance strategies. Compared to periodic-based approaches it stands out due to the reduced maintenance effort as well as due to the omission of unexpected machinery downtime. The continuous supervision and diagnosis of equipment for capturing actual states, also known as condition monitoring, is a prerequisite of condition-based maintenance. Condition monitoring can be conducted manually by acoustic or visual inspection or automatically by condition monitoring systems collecting relevant data, such as temperature and force, with computational sensors.[49] Condition-based maintenance is based on the assumption of looming failures. This means that critical operating conditions don't occur abruptly and escalating or with other words, failures loom in the behavior of state variables collected by the monitoring system. As a consequence, anomalies are detected prior to breakdown and can be avoided by timely and appropriate measures.

As shown in Figure 6, the time span between failure detection and occurrence is called PF-Interval and represents the advance warning time. The bigger the interval the more time is left for taking corrective measures and preventing impending breakdown. However, in general the PF interval is unknown and may vary between milliseconds and years depending on the component and failure type. Therefore, the main effort in condition-based maintenance focusses on developing methods for improving early recognition and increasing the PF interval.[50]

[47] cf. Fitch, 1992, p.13
[48] cf. Ansari/Glawar, 2019, pp.320f
[49] cf. Reichel/Müller/Haeffs, 2018, 354f
[50] cf. Matyas, 2019, p.125

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

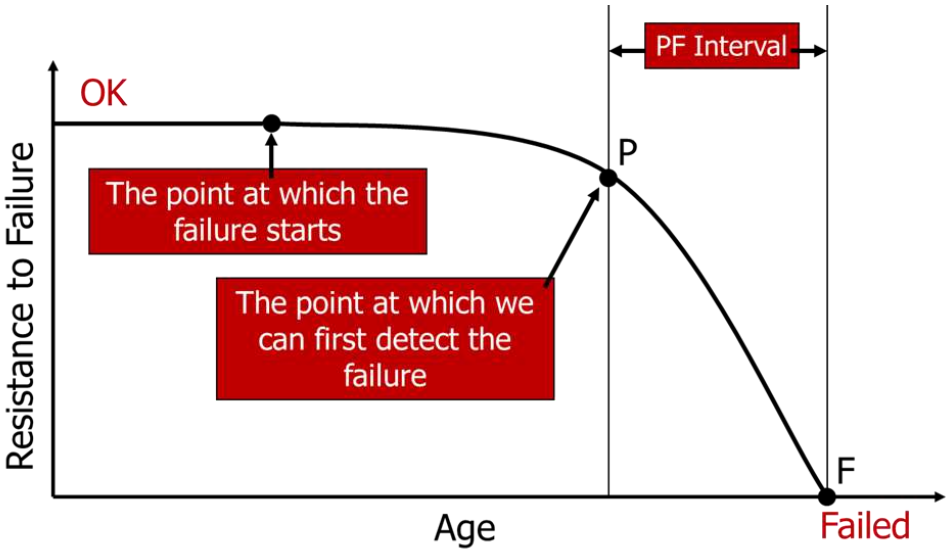

Figure 6: PF Interval of component lifecycle[51]

As the condition-based strategy, **predictive maintenance** is based on the state of plant equipment using monitoring systems or inspection methods for data collection. For obtaining the actual operating condition, vibration monitoring, thermography and tribology have turned out to be the effective tools and consequently are commonly used in industrial applications.[52] However, compared to the condition-based approach, predictive maintenance is not only limited to the detection of impending failures and break downs. Instead of relying on obvious looming of critical operating conditions in actual plant data, it applies advanced statistics, stochastics, real time analytics or even machine learning algorithms for predicting the time of failure occurrence. Therefore, the enhancement of condition monitoring by advanced statistics and algorithms does not only improve the disclosure of looming failures, it reveals patterns and correlations for predicting the remaining lifetime of objects, which is crucial for planning and scheduling in maintenance. Referring again to Figure 6, predictive maintenance investigates the problem of inconstant and variable PF-intervals and provides forecasts of the remaining useful life of objects (RUL). There are three main approaches for predicting the future behavior of systems: Physics-based, data-driven and model-based techniques. Physics-based approaches are characterized by experiments for determining parameters of formulas and correspond to the class of empirical methods. In contrast, data-driven techniques are solely based on historical data and investigate the correlation between plant equipment's current health condition and future state. This includes artificial intelligence algorithms, supervised and unsupervised machine-learning methods such as recurrent neural networks and cluster algorithms, but also statistical analysis using probability distributions such as the Weibull-distribution. Finally, the model based prognostic method combines the two previous described approaches for

[51] URL: https://www.assetivity.com.au/article/reliability-improvement, 25.05.2019
[52] cf. Mobley, 2002, pp.4f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

increasing accuracy and robustness. It extends observations and experiments with probability distributions and intelligent algorithms for including uncertainty and predicting future events and states. Important representatives of that category are Particle Filters also known as Sequential Monte Carlo methods as well as Kalman filters.[53]

Due to the increasing complexity of production processes, the classical strategies, reactive-, preventive-, proactive- and condition-based maintenance, often provide insufficient solutions for guaranteeing required equipment availability. Especially in flexible and customer-oriented manufacturing systems with varying production programs and load spectrum, these conventional strategies are stretched to their limits due to the changing scope conditions. Therefore, a predictive, anticipative and comprehensive maintenance strategy, which includes sensor data from condition monitoring systems, quality and machinery data as well as knowledge of historical failure events, is required.[54] Investigating applied predictive maintenance models in real production systems, approaches mainly rely on the analysis of machinery states striving for failure detection and predicting remaining useful life. Indeed, forecasting failures is an essential part for preventing unexpected downtime. Nevertheless, analyzing difficulties in industrial maintenance, inappropriate strategies for planning and monitoring as well as management mistakes have turned out to be the main causes of troubles in maintenance.
In order to overcome these difficulties, **prescriptive maintenance** deals with the gap of missing comprehensive knowledge providing decision support and recommendation systems for enhancing the predictive strategy to the next level. It does not only answer the question "What will happen when?" and predict the time of failure occurrence, it is dedicated to the problem area of decision making and answers the question "How should a specific event happen?"[55] Thus, prescriptive maintenance means shifting from conventional preventive strategies to proactive smart maintenance planning, recommending course of actions and maintenance measures.[56] The main goals include speeding up the decision making processes, improving maintenance workflow, optimizing impending maintenance actions and finally reducing the total costs of manufacturing. Prescriptive maintenance is the most mature approach known so far, which also includes complex methods for developing and reinforcing adaption and optimization capabilities. It is also seen as an essential approach for achieving self-organizing systems, which enable self-diagnosis and automated maintenance scheduling.[57]

[53] cf. Gao/Wang/Teti/Dornfeld/Kumara/Mori/Helu, 2015, p.750
[54] cf. Matyas, 2019, pp.136f
[55] cf. Ansari/Glawar/Nemeth, 2019, p.1
[56] cf. Matyas/Nemeth/Kovacs/Glawar, 2017, p.461
[57] cf. Matyas, 2019, pp.332f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Realizing prescriptive maintenance approaches requires a comprehensive maintenance knowledge base, which has to be extracted from heterogeneous, multimodal and multidimensional data sources. The exploitation of that maintenance knowledge is characterized by searching, providing, retrieving and using existing knowledge which also includes domain expert's background knowledge gained by years and years of work. Apart from that, decision recommendation or even automated decision-making prerequisites the disclosure of new knowledge, also referred to as knowledge exploration. Its goal is to gain new insights and find new inferences for extending the existing strategies. Thus, prescriptive maintenance aims at enriching and preserving existing knowledge but also exploring new knowledge for enhancing decision making based on multiple knowledge sources.[58]

One arising challenge of prescriptive models is the collection of relevant data and information and the creation of a comprehensive data base. Depending on duties, individual preferences and alignment of maintenance strategies towards quality- or load orientation, existing approaches tend to either associate production planning and maintenance data or to combine product quality and maintenance planning. In contrast to that, novel prescriptive maintenance approaches try to benefit from the consolidation of all mentioned sources with respect to process, product and machine aspects.[59] Figure 7 illustrates these data categories and their different sources with a focus on equipment and human interaction. Notably, the data types depend on the respective source. Human communication and documentation are characterized by unstructured data such as texts and speech, whereas equipment data is typically structured.

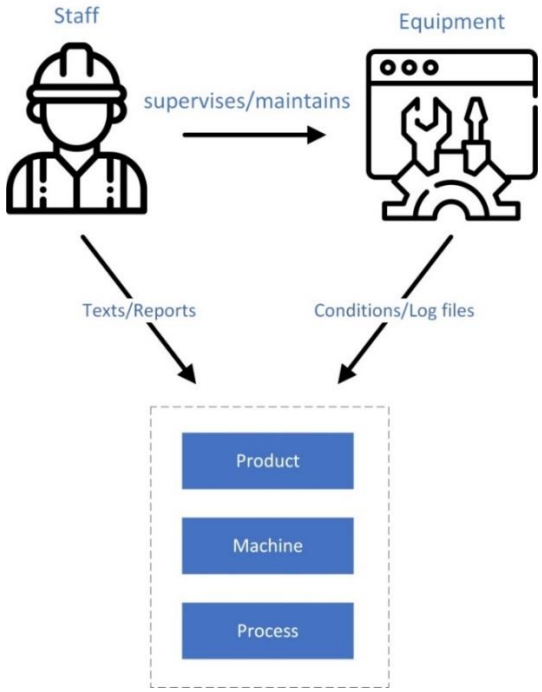

Figure 7: Different data sources and categories

[58] cf. Ansari/Glawar/Nemeth, 2019, p.2
[59] cf. Matyas/Nemeth/Kovacs/Glawar, 2017, p.461

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Due to the multidimensionality of data sources and versatility of objectives, prescriptive maintenance models are complex constructs being structured in different layers. An appropriate foundation for the architecture of prescriptive maintenance models is the Cross Industry Standard Process for Data Mining (CRISP-DM), consisting of six phases as follows.[60]

- *Business understanding:* It's the first phase and focuses on becoming acquainted with the project and understanding its difficulties, objectives and requirements from the business point of view.
- *Data understanding:* This layer is characterized by the first inspection and assessment of the data for realizing data quality problems and detecting interesting subsets in order to form hypotheses.
- *Data preparation:* This phase includes all measures required to build the ultimate dataset from the initial raw data, such as attribute selection, data preprocessing and transformation as well as data cleaning for modeling tools.
- *Modeling:* The modeling methods are selected and applied with respects to objectives and requirements. Some modeling techniques prerequisite specific preprocessed data, which could lead to an iteratively pass of Data preparation phase.
- *Evaluation:* The evaluation phase assesses the quality of the model and determines whether all relevant business objectives are met.
- *Deployment:* The implementation and deployment of the model are planned, and further steps are recommended to the customer in order to ensure a smooth and trouble-free introduction phase.[61]

2.1.3 Knowledge-based Maintenance

Knowledge-based maintenance is a key enabler of digital transformation and reinforces the shift from conventional strategies to predictive and prescriptive maintenance. This new generation of strategies acts or even decides based on maintenance knowledge and goes beyond the assessment of deterioration and conditional state.[62] Knowledge-based maintenance starts from the premise that competitive advantages can't be accomplished by traditional approaches for cost reduction. Especially in environments of high personnel costs and taxation, the scope of action of conventional strategies is limited to optimizing performance by improving efficiency and quality. On the contrary to that, knowledge-based maintenance enables a holistic view of the whole system and considers long term effects of strategical decisions. Instead of focusing on sub-domains, it aims at comprehensive optimization, allowing the discovery of unrevealed potential and increasing the impact

[60] cf. Ansari/Glawar/Nemeth, 2019, p.13
[61] cf. Chapman/Clinton/Kerber/Kahabaza/Reinartz/Shearer/Wirth, 2000, pp.12f
[62] cf. Ansari/Glawar, 2019, p.318

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

on overall costs. The main knowledge sources are provided data in EDP systems but also domain-knowledge from employees and managers.[63]

Figure 8 gives an overview of knowledge-based maintenance and should emphasis the holistic idea behind this approach. The first row illustrates the main knowledge base components: System conditions, maintenance consequences and management processes, which altogether represent the three organizational levels: Strategic, tactical and operational levels. The gathered machine, process and product data are then provided to the overall maintenance strategies: risk-based, condition- or time-based and Total Productive Maintenance.[64] The mentioned strategies and abbreviations are defined as follows: Condition-and time based maintenance have already been dealt with in section 2.1.2 and rely on the actual state of an object or respectively are based on predefined time intervals. Reliability-centred maintenance (RCM) is a structured guideline for determining an appropriate preventive strategy suggesting maintenance measures and intervals.[65] The consideration of environmental impacts and preservation of safety as well as issues of reliability are combined in the risk-based maintenance approach which is commonly applied in the area of maintenance investing.[66] Total Productive Maintenance (TPM) includes plant equipment's entire life span and involves every single employee aiming at the continuously improvement of quality and productivity.[67] What's more Total Life-Cycle Cost (TLC) pursues a comprehensive life cycle approach including initial costs and maintenance costs for inspection, repair and improvement[68] Finally, Lean Maintenance applies the principles of lean production to industrial maintenance and strives for identifying and eliminating non-value adding maintenance activities.[69] Knowledge-based maintenance is the superior instance of all mentioned areas and strategies and considers approaches of all different layers.

[63] cf. Pawellek, 2013, p.8
[64] cf. Ansari/Glawar/Nemeth, 2019, p.4
[65] cf. Kobbacy/Murthy, 2008, p.79
[66] cf. Khan/Haddara, 2003, p.561
[67] cf. Chan/Lau/Ip/Chan/Kong, 2005, p.72
[68] cf. Frangopol/Lin/Estes, 1997, p.1
[69] cf. Mostafa/Dumrak/Soltan, 2015, p.434

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

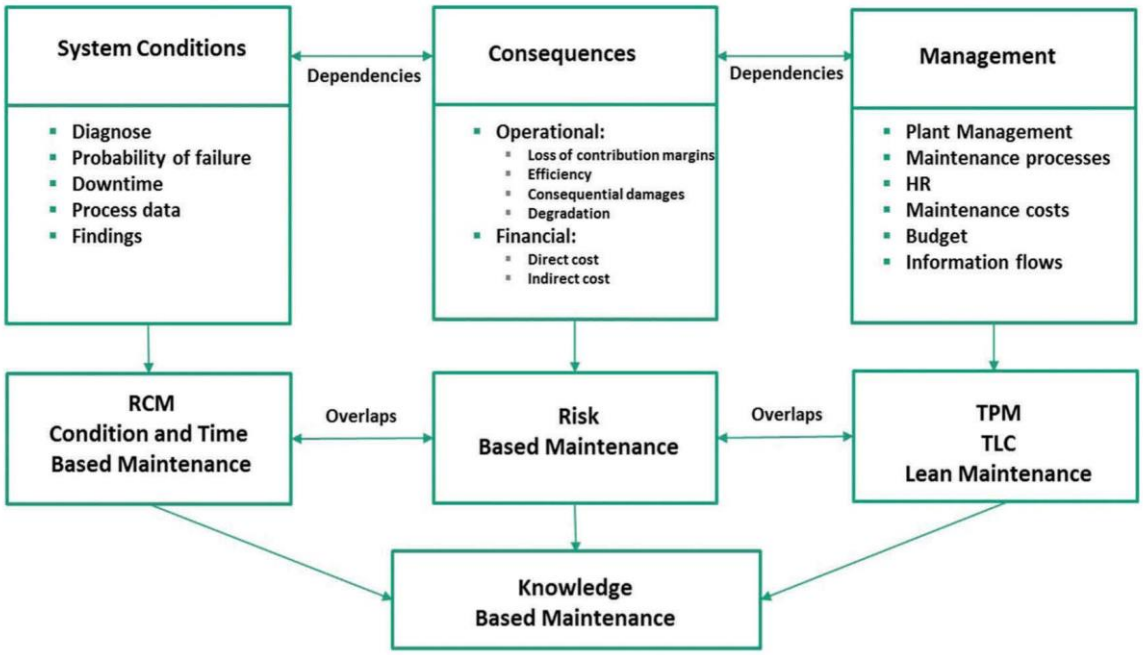

Figure 8: The basic model of knowledge-based maintenance[70]

Based on that holistic approach maintenance strategies evolved in course of time which can be categorized in different maturity and complexity levels. The most common strategies and concepts have already been described in chapter 2.1.2 and are not discussed in detail again. Nevertheless, the following Figure 9 should help to structure the general approach of knowledge-based maintenance and additionally provide a classification of strategies from a Knowledge Management perspective.

Representing the lowest level of KBM, Descriptive Maintenance focuses on gathering information in order to depict failure occurrences, characterize events and visualize maintenance operations. In Diagnostic Maintenance that approach is extended by applying statistical analysis for obtaining cause-effect relations and for providing further information about the reasons of unexpected happenings which is referred to as Knowledge Generation. The next stage deals with the generation of forecasts predicting future events and failures using historical and real time data. It is characterized by the application of advanced analytical and statistical methods such as supervised and unsupervised machine learning for enabling knowledge discovery. Finally, Prescriptive Maintenance represents the most advanced most sophisticated level of KBM, recommending or even self-deciding future maintenance operations for global optimization. It requires dynamic capabilities for adapting to new circumstances and providing proper suggestions.

In the approach illustrated in Figure 9 supposed by Ansari et. al Predictive and Prescriptive maintenance are linked to a feedback loop which evaluates former prediction for improving future recommendations. The resulting interaction of these

[70] Ansari/Glawar/Nemeth, 2019, p.8

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

traditionally separately regarded strategies allows exploiting synergies between prediction and prescription achieving improvements in prediction accuracy as well as in action planning and recommendation quality. The consolidation covers fields and tasks such as:

- modeling and representing domain specific knowledge:
- predictive data analytics
- semantic reasoning[71]

Domain specific knowledge is a closed sphere having its own knowledge repository such as individual dictionaries for abbreviations and domain specific terms. Common approaches for representing domain-specific knowledge are ontologies, which describes relations and rules of vocabulary for combining terms.[72] Predictive data analytics involve all kinds of techniques aiming at predicting future states being classified in Physics-based, data-driven and model-based methods.[73] Finally, semantic reasoning applies rules to the created ontology and derives new knowledge from the represented properties, relations and constraints.[74]

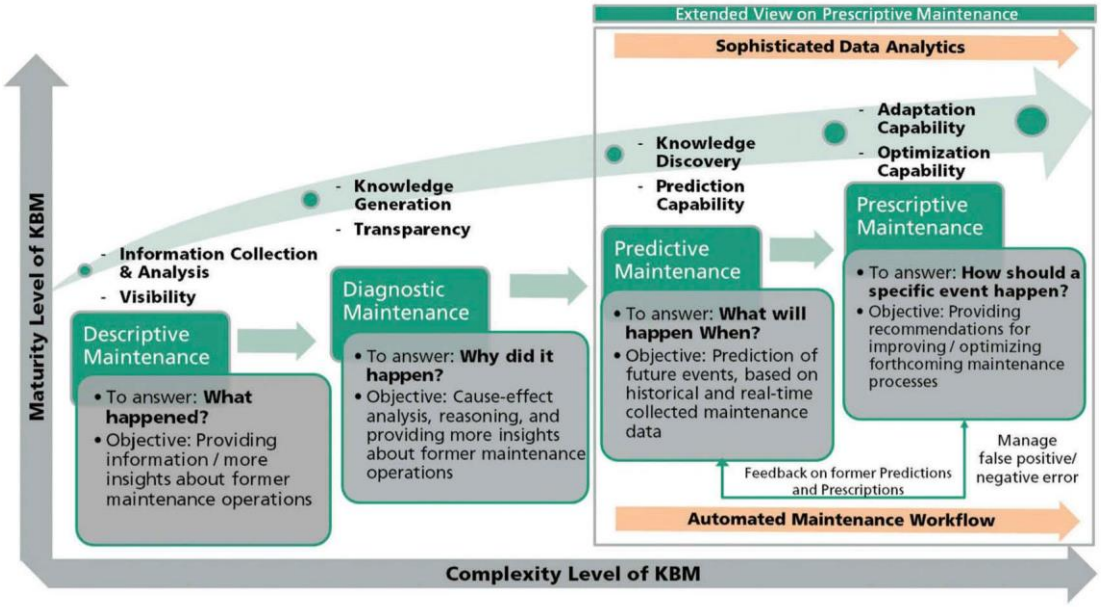

Figure 9: Knowledge-based Maintenance evolution[75]

Taking into account the broad variety and interaction of strategies, IT-systems, organizations and data consolidated in the approach of KBM, multiple challenges and problem areas arise which require proper and sophisticated solutions. The problem-solving process is strongly affected by the multilateral complexity of KBM and can be

[71] cf. Ansari/Glawar/Nemeth, 2019, pp.4f
[72] cf. Yoon/Dankel, 2005, p.1
[73] cf. Gao/Wang/Teti/Dornfeld/Kumara/Mori/Helu, 2015, p.750
[74] cf. Karray/Chebbel-Morello/Zerhouni, 2011, pp.3f
[75] Ansari/Glawar/Nemeth, 2019, p.9

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

symbolically compared with a Rubik's cube. Each cell of the cube represents a dynamic and data-driven unit which influences the process of problem solving. Overcoming the complexity of a Rubik's cube requires the recognition of patterns but is also driven by Trial and Error procedures. Transferred to the challenges of KBM, well-structured problem solving and identification as well as learning and reasoning from historical events based on logical capabilities are the key elements for mastering the arising difficulties. The mentioned multilateral complexity can be described with the following three dimensions:[76]

- *Multidimensionality of maintenance organization, stakeholders and IT-Systems:* The organization of industrial maintenance is structured in operational, tactical and strategic levels impersonated by maintenance operators, engineers, project managers and top-managers with different duties and perspectives. Moreover, the IT-landscape usually consists of various systems being linked by interfaces and providing machinery-, process-, resource-, quality and cost data as well as Key Performance Indicators.
- *Multi-modality of data:* Linking of different sources of data with different information content provides independent information and inferences of maintenance aspects. For example, a failure signal could indicate a machinery component break down, but also be traced back to inappropriate planning causing deterioration of sub-components.
- *Different maintenance strategies and approaches:* KBM includes strategies providing standards and recommendations for industrial maintenance implementation and operations such as TPM, TLC or RCM. What's more, approaches without sensing technologies like preventive strategies but also computational-technology-based strategies such as prescriptive maintenance are involved.[77]

2.2 Text-Mining

“Text-mining is the discovery and extraction of interesting, non-trivial knowledge from free or unstructured text.”[78] For clarifying the logic and the meaning behind the introduced definition, it is purposeful to discuss the seemingly trivial but sometimes misleading terms “data”, “information”, “knowledge” and explain the different types of data first.

Data are recognizable changes of physical states varying in space, time and energy which are stored and registered. Finding significant coherences and patterns in data,

[76] cf. Ansari/Glawar, 2019, p.319
[77] cf. Ibid., pp.319f
[78] Kao/Poteet, 2007, p.1

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

results in information, whereas the determination of the significance level is up to the analyst. This means, information acquired from data is subjective and depends on the person extracting it with its individual expectations and hypotheses. The extracted information has then an impact on hypotheses and expectations held by the person and can trigger the generation of new hypotheses. All gathered hypotheses of a person are now referred to as knowledge and the creation of new hypotheses out of information is defined as knowledge discovery being situated in a mutual dependency and interaction field.[79]

Although the classification of information and data might appear inappropriate due to the subjectivity of information, the separation in three different categories driven by the computational accessibility turned out to be practicable.[80] To begin with, structured data represents the easily computable class and is characterized by its regular format and predictability. Typically, structured data can be illustrated in tables containing records, attributes, keys and indexes and is mostly managed by a database management system based on a data model. As a result, specific information can be found quickly and easily.

On the opposite, unstructured data is- due to its lack in regularity- difficult to process and unrecognizable for computers. Locating specific data units requires sequential search in long data strings being also referred to as parsing and comes along with an increased run-time compared to structured environments.[81] Moreover, the extracted information is strongly linked with the underlying data and cannot be fully separated from the data itself. The most common representative of unstructured data is text. The meaning of text depends on the text itself, which means on the arrangement of words and on the context. Separating words from its original arrangement could change the meaning of a text. Other common examples of unstructured data are audio or video files.[82]

Besides, semi-structured data is a mixture of the previously introduced types, underlying a general structure but being partially not separable from the data itself.[83] For example, e-mails are structured in sender, recipient, subject and message, but the text itself is unstructured. Moreover, textual formats based on languages such as Extensible Markup Language (XML) belong to the semi-structured class providing rules and structures for data representation.[84]

[79] cf. Holzinger/Stocker/Ofner/Prohaska/Brabenetz/Hoffmann-Wellenhof, 2013, p.15
[80] cf. Ibid., p.15
[81] cf. Inmon, 2015, p.9
[82] cf. Holzinger/Stocker/Ofner/Prohaska/Brabenetz/Hoffmann-Wellenhof, 2013, pp.15f
[83] cf. Buneman, 1997, p.1
[84] cf. Chen/Wang/Liu/Lin, 2009, p.1005

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

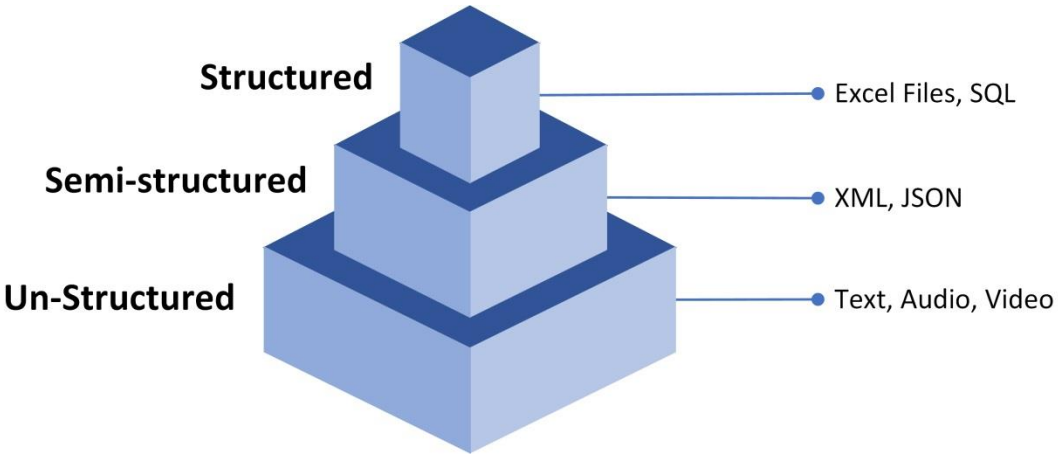

Figure 10: Examples of structured-, semi-structured and unstructured data[85]

In literature of Human Language Technology, text-mining approaches for language understanding, speech recognition and language generation are also referred to as Natural Language Processing (NLP).[86] NLP imitates the human ability of text comprehension and involves algorithms and methods for deciphering and recognizing ambiguities in human language.[87] The usage of the terms text-mining and NLP is not always clear and consistent, whereas text-mining and NLP are also sometimes regarded as separate branches. In this work the term text-mining also refers to algorithms and methods of NLP which both are considered as part of the general topic of artificial intelligence (AI). Another expression which commonly appears in related literature and might lead to confusion is Knowledge Discovery from Text (KDT). Since KDT and text-mining pursue the same goal of extracting knowledge from textual data, they can be put on the same level and be used interchangeably.

Accomplishing knowledge extraction from text requires the explication of latent information and the extraction of semantic relations, which is essential for comprehending unstructured natural language. This involves, for example, the recognition of entities or the annotation of part-of-speech in text, which is a negligible task for humans, but demands sophisticated models and algorithms for computational analysis. Therefore, text-mining draws on methods from machine learning, information extraction and knowledge management in order to master the challenging task of text comprehension.[88] Despite the necessary efforts for imitating human abilities, computational analysis stands out by its capability of processing huge amount of data which could not be handled by manual evaluation any more. According to Yu et al. text represents 80 percent of corporal information building a big

[85] cf. URL: https://www.phaedrasolutions.com/structured-data-vs-unstructured-data/, 26.07.2019
[86] cf. Kao/Poteet, 2005, p.1
[87] cf. Martinez, 2010, p.352
[88] cf. Gupta/Lehal, 2009, p.60

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

data pool for computational processing.[89] Apart from the angle of data amount and availability, text contains valuable latent knowledge, such as opinions and object descriptions, which businesses can benefit from.[90] For instance, the extraction of sentiments in numerous product reviews allows companies to evaluate their products and provides important feedback for further decisions.[91] Moreover, knowledge discovery and exploitation is used for managing and conducting maintenance activities, which can be supported by the analysis of text documents.[92] Thus, text-mining as a sub-domain of data-mining has been detected as potential tool for decision support and knowledge extraction being implemented in various sectors as discussed in more detail in Section 3.[93]

A general framework for the processes of Text-mining is shown in Figure 11 being split up in three stages and listing the most common methods of the corresponding category. The methods and approaches are described in detail in the following sections. Furthermore, their application is demonstrated in Chapter 5.

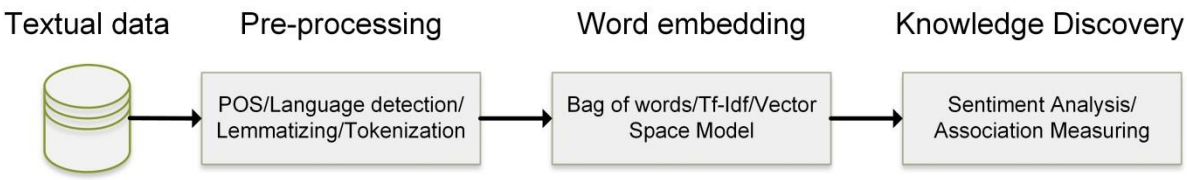

Figure 11: General framework of Text-mining[94]

2.2.1 Text Pre-processing

Text pre-processing involves all procedures and methods for cleansing tokens and preparing the text for analysis. Tokens are defined as single words or sets of words representing the smallest unit in texts. Pre-processing tokens includes tasks such as correcting spelling mistakes, removing words with low information content and transforming words to their root-form. It is individually adapted to the text-mining application with a focus on the retention of relevant information content.[95]

a.) Tokenizing

The structure of text has many levels. Documents are composed of paragraphs; paragraphs consist of sentences and sentences are sequences of tokens. A token as the smallest unit for morphological processing is often a word but could also

[89] cf. Yu/Wand/Lai, 2005, p.35
[90] cf. Pathak, 2014, p.137
[91] cf. Aggarwal/Zhai, 2012, p.416
[92] cf. Ansari/Uhr/Fathi, 2014, p.15
[93] cf. Dörre/Gerstl/Seiffert/Fayyad, 1999, p.400
[94] vgl. Aggarwal/Zhai, 2012, p.389
[95] cf. Ibid., p.389

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

represent a combination of words, which regularly appears in the German language. For example, the term “zum” is in fact a mixture of two words “zu” and “dem”, but “zum” is recognized as a single token. Therefore, sequencing tokens enables the reconstruction of the original text, but sequencing words could lead to deviations.[96]

Tokenization is now defined as the procedure of transforming text into a sequence of tokens.[97] The necessary effort for accomplishing word tokenization depends strongly on the underlying language. Texts for example of the English or German language are split according to delimiters or sets of delimiters, such as space or punctuation symbols, which enormously facilitates word segmentation. In contrast to that, Chinese words are not delimited at all, which turns tokenization to a serious obstacle.

On the larger scale of sentence tokenization tasks show up to be more challenging even for Latin alphabet-based languages because of the ambiguity of the most common delimiter, the period. Periods indicate the end of a sentence, but also mark abbreviations. Therefore, additional preprocessing steps like abbreviation expansion are required in order to obtain satisfying sentence segmentation.[98]

Common tokenizers, such as the PTBTokenizer from Stanford NLP group, utilize lexical analyzer generators.[99] A lexical analyzer generator takes as input a language specification being composed of regular expressions and corresponding actions. It generates a program that reads input, matches the input against the regular expressions in the language specification file, and runs the corresponding action if a regular expression matched. Thus, the input sequence of characters is converted to a token stream based on the underlying language specification. Mathematically the procedure is based on Deterministic Finite Automation (DFA), which is characterized by a sequence of finite number of states. Having a string of characters as input, the DFA checks state transaction rules which define whether the string may proceed to the next state. In the end the final state position reveals whether the input string is accepted or rejected.[100] The following figure shows a simple example of tokenization in German language taken from the data base of maintenance records analyzed in the use-case.

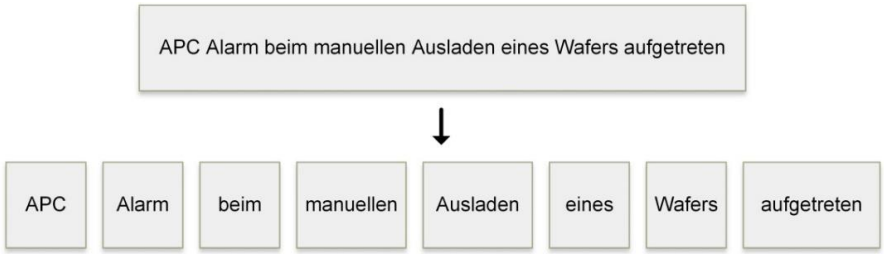

Figure 12: Example of tokenization

[96] cf. Straka/Strakova, 2017, p.89
[97] cf. Holzinger/Stocker/Ofner/Prohaska/Brabenetz/Hoffmann-Wellenhof, 2013, p.20
[98] cf. Benesty/Sondhi/Huang, 2008, p.459
[99] cf. URL: https://nlp.stanford.edu/software/tokenizer, 14.11.2019
[100] cf. URL: https://jflex.de/, 14.11.2019

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

b.) Stemming

After text has been split into a sequence of tokens, the next purposeful step is to convert tokens to their root form, which is usually referred to as stemming. Its goal is to standardize words and eliminate grammatical variations of words depending on language specific conventions such as conjugations, declinations and the building of participles. Stemming is a highly recommended preprocessing step especially for algorithms based on word frequencies, since grammatically adapted words occurring in different inflections would otherwise be processed as individual instances.[101] There are different types of stemming approaches, of which the most common one is the Porter Stemmer for the English language. On a general perspective it is based on five steps for word reduction, whereas each carries conventions for rule selection. An important convention is for example the selection of rules which apply to the longest suffix, which is also shown in the table beneath:

Table 1: Rules and examples of Porter Stemmer[102]

Rule	Example
SSES → SS	caresses → caress
IES → I	ponies → poni
SS → SS	caress → caress
S →	cats → cat

Many of the following conventions and rules are based on the number of syllables of a word in order to avoid over-stemming. Considering the rule and example in the following table, the word replacement exceeds the required number of syllables m and the rule is applied, whereas short words such as cement are not affected by the rule.[103]

Table 2: Measure based rule for stemming[104]

Rule	Example
(m > 1) EMENT →	Replacement → replac

Of course, the Porter algorithm covers a huge set of intricate rules and conventions which are not explained in detail in this work. The introduced examples and corresponding rules however should have indicated its general principle and the underlying methodology.

Figure 13 reveals the problems which could arise by using stemming and provides three representative examples. The first one correctly converts the gerund form to the initial verb by detecting and chopping off the suffix. However, as shown in the second

[101] cf. Weiss/Indurkhya/Zhang, 2015, p.19
[102] cf. URL: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization, 14.11.2019
[103] cf. Ibid.
[104] cf. Ibid.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

line, removing the suffix could also result in inappropriate terms depending on the morphology, or with respect to the presented example, depending on whether the verb is used with the third person or it's a noun using the gerund. Especially in the German language obtained stems might be unsatisfying because of the more complex building of tenses and participles which is illustrated in the third example.

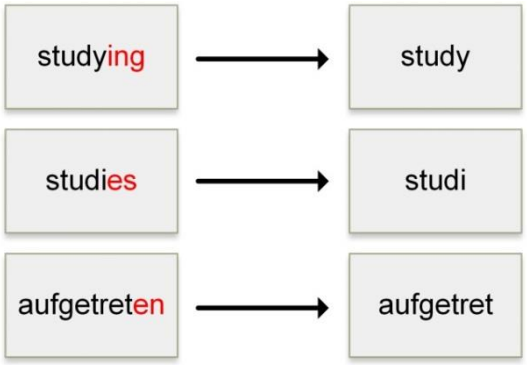

Figure 13: Examples of stemming

c.) Lemmatization

Lemmatization is similar to stemming and pursues the same objective of finding the root form of words in order to reduce the number of distinct tokens. As the procedure of stemming, it relies on the premise that words with equal roots share the same meaning. Nevertheless, lemmatization is a more complex approach considering additional morphological information. It provides solutions for words having multiple meanings and detects root forms not only due to the appearing of pre- and suffixes, but also due to the grammatical structure. That morphological information involves attributes such as tense, person, number and part-of-speech. Especially in morphological rich languages lemmatization achieves higher accuracies justifying the increasing complexity compared to stemming.[105]

Common algorithms for lemmatization include rule-based approaches being characterized by the consistent addition of rules to an existing set of rules and supervised machine learning approaches trained on sets containing suffixes and class labels.

Rule-based approaches require a system built up upon the most general rules. For example, in English a rule could be: "ed" is added to verbs for forming the past tense. To these fundamental if-then-else statements except-branches are added in case of incorrect conclusions or else branches if no rule fires. That procedure ensures rule structure extension with consistency and provides a decision tree with a tailored solution for the problem of lemmatization. The morphological information such as

[105] cf. Anandarajan/Hill/Nolan, 2019, p.54

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

POS-tags and word-suffixes represent conditions in the rule statements and the obtained output the root forms of the original words.[106]

Lemmatization by supervised machine learning is a more advanced approach based on a labeled training set searching for patterns and regularities. The term "labeled" indicates that the training data consists of input vectors together with their corresponding output value. Applied to the field of lemmatization an input vector could for example represent the characters of a word and the output value the corresponding lemma. Based on that information supervised machine learning strives for creating a function reconstructing the classification results of the labeled data set by iteratively adapting the parameters of the function. Therefore, the algorithm learns from the labeled training data set and derives a function which finally enables the classification of new unlabeled examples.[107]

A novel and successful supervised machine learning approach for accomplishing lemmatization is sequential modeling with recurrent neural networks. Since neural networks are an important method occurring in all different layers of text-mining, the following section briefly introduces the general principle and then focusses on sequential modeling in combination with lemmatization.

Artificial Neural Networks (ANN) are inspired by the human brain consisting of neurons which are able to communicate by transmitting electrical signals with the help of synapses. The biological neural structure is modelled by a network with each node being able to receive, process and transmit signals. Figure 12 illustrates a single node referred to as artificial neuron with three input values defining the input vector $= (X_0, X_1, X_2, X_3)$, whereas $X_0 = 1$ because of mathematical reasons which is explained later on. The synaptic weights $\boldsymbol{W} = (w_0, w_1, w_2, w_3)$ act as controller of transmission properties and are responsible for the enhancement or weakening of received signals through multiplication with the corresponding input values. Building the sum of the products results in a signal $v = \sum_{i=0}^{3} W_i * X_i$. also referred to as total sum of postsynaptic potential. Finally, it is checked whether the postsynaptic potential exceeded a specific threshold and the neuron is activated resulting in an output value equal one. This activation behavior is modelled by the activation function f which is shown in Figure 15 and called Heaviside step function. It is defined as follows:

$$y(v) = \begin{cases} 0, & v < 0 \\ 1, & v \geq 0 \end{cases}$$

Equation 1: Heaviside step function

which means the neuron fires if the postsynaptic potential v exceeds zero. Since the input value $X_0 = 1$ the threshold can be influenced by the weight w_0 being added up

[106] cf. Plisson/Lavrac/Mladenic, 2004, p.2
[107] cf. Ertel, 2017, p.177

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

with the weighted input values. For $w_0 = 0$ the threshold equals zero.[108] Notably, there are also other forms of activation or transfer functions depending on the type of network and application, such as sigmoid or rectifier function.

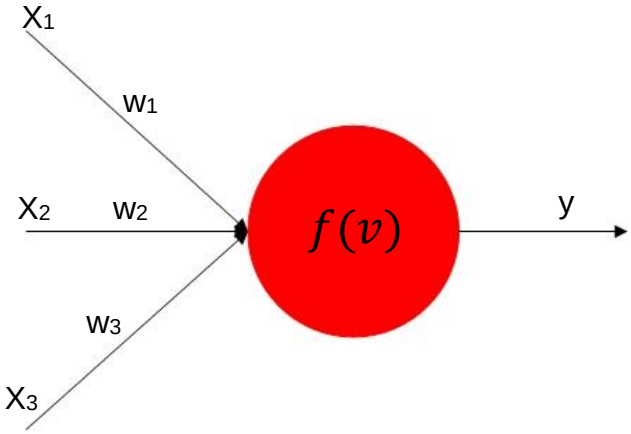

Figure 14: Model of artificial neuron

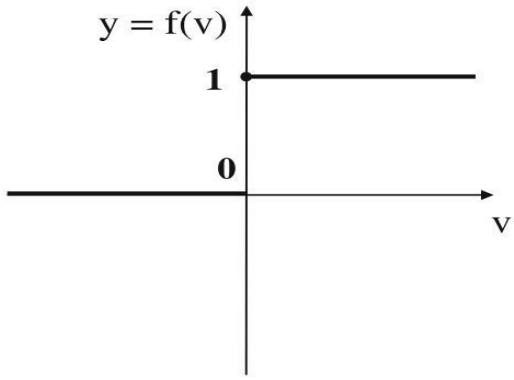

Figure 15: Heaviside step function[109]

Based on the introduced concept demonstrated on a single neuron, networks are formed by linking the neurons according to specific rules generating an extremely simplified model of the human brain. As illustrated in Figure 16, ANN consist of different layers, input, hidden and output layers being built up by several neurons illustrated as colored nodes. Each neuron is connected with every neuron of the following layer, however, there is no interaction with nodes of the same layer. Thus, the output value of a single neuron depends on the values received from the neurons in the previous layer. This leads to the following equation for the output value of a neuron, whereas j is the considered neuron and n the number of neurons in the previous layer.[110]

$$x_j = f(\sum_{i=1}^{n} w_{ji} * x_i)$$

Equation 2: Output value calculation of an artificial neuron

The size and architecture strongly depend on the application and the available training data. Especially the depth and width or with other words, the amount of hidden layers and the number of neurons within each layer have to be adjusted according to the task and required results. The number of nodes in the input layer depends on the available data and the amount of features representing the input objects. For example, in machine vision each pixel of an image could be an input value. The dimension of the output layer is designed based on the required

[108] cf. Flasiński, 2016, pp.158f
[109] Ibid., p.160
[110] cf. Ertel, 2017, p.247

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

classification. For example, the classification of images in three categories implies three nodes in the output layer.[111]

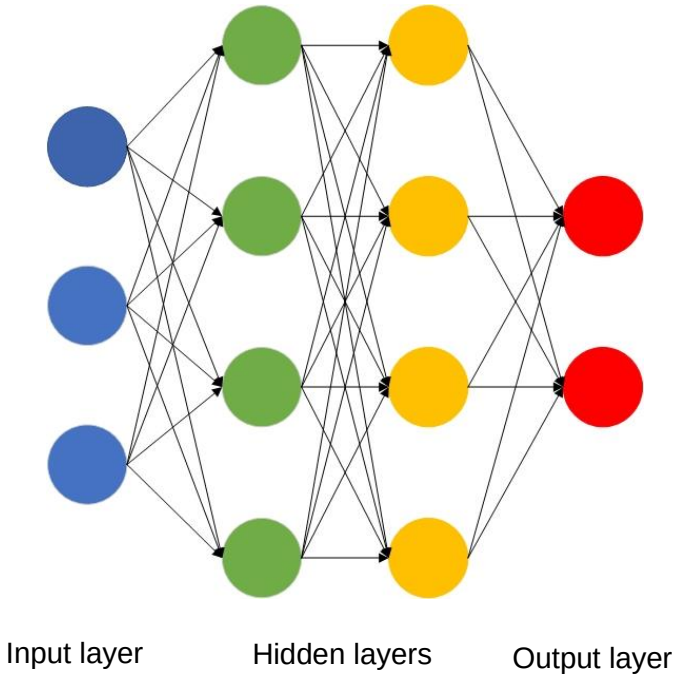

Figure 16: ANN with multiple layers

The fundamental task in supervised machine learning is the knowledge extraction of the provided labeled training data and the determination of the optimal mapping function for obtaining the required outputs. The training data consists of vector pairs, being composed of an input vector and a corresponding output vector. The mapping function is represented by the thousands or even millions of weights of the neural network being responsible for the correct processing of the input values. Therefore, the main target is the determination of optimal weight values for reconstructing the results given in the labelled data set, which is also referred to as neural network training.[112]

There are several training algorithms for adjusting the weights of artificial neural networks which differ in performance, speed and precision. Nevertheless, all of them pursue the same goal of efficiently solving a minimization problem being quantified by a loss function. For getting an overview and explaining the major steps of the procedure the Backpropagation algorithm, which counts to the most used, is briefly introduced.

It starts by the random selection of weights and the calculation of the associated output for one training data item based on that selection. The quadratic error of the obtained output vector and the target output vector gives the loss function for this single training example which serves as an indicator for the current weight selection.

[111] cf. Ertel, 2017, p.245
[112] cf. Ibid., pp.179f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Building the partial derivatives after the weights of this function reveals information about whether weights have to be increased or decreased and shows the weight sensitivity towards the individual loss function. The procedure is repeated until the loss functions as well as the corresponding derivatives for all training examples are calculated. Since the calculus is accomplished recursively starting from the hidden layer and moving backward to the first layer, the algorithm is referred to as Backpropagation. In the next step the average for each weight is built over the partial derivatives being formed to the negative gradient of the total loss function. Iteratively adapting the weights proportional to the gradient and recalculation finally leads to the minimum of loss and the optimal weight configuration.[113]

Coming back to the tasks of text pre-processing, supervised machine learning and, in particular, neural networks have turned out to be a useful tool for including the morphological aspects which, for example, are essential in the prediction of lemmas. However, in the introduced concept of neural networks the input and output vectors are of constant length which is an unpractical attribute in text processing. For instance, analysing whole sentences appears unsuitable because of the arbitrary number of words but also the processing of single tokens is unpractical due to the differences in length. Therefore, neural network lemmatization draws on recurrent neural networks which enable the processing of input sequences.[114] Recurrent neural networks stand out by providing a memory which stores information about previous computations. Thus, the current state of a neuron can also depend on previous inputs enabling sequential processing. As shown in Figure 17, one possibility of implementation are feedback-loops storing the state information and passing it back as input value for the same neuron, whereas the black square indicates the state memory of the neuron. The following output calculation equals the already mentioned procedure. There are many different types of recurrent neural networks providing distinct solutions and architectures for implementing dynamic behavior and sequential processing. This work focusses on the general principle and will leave the coverage of various approaches open for further discussion.[115]

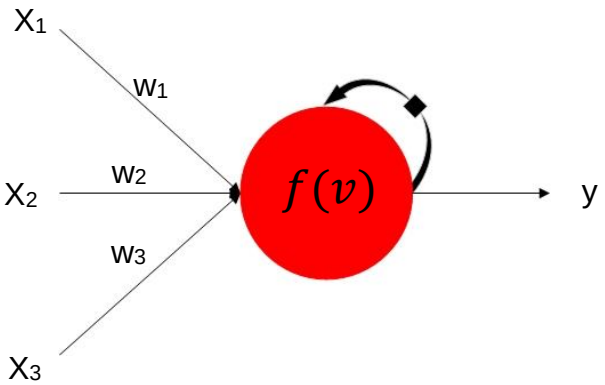

Figure 17: Recurrent neuron

[113] cf. Ibid., pp.269f
[114] cf. Cho/van Merrienboer/Gulcehre/Bahdanau/Bougares/Schwenk/Bengio, 2014, p.3
[115] cf. Flasiński, 2016, pp.170f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Novel approaches for accomplishing neural network lemmatization use sequence to sequence architectures consisting of two consecutively arranged recurrent neural networks of which the first is referred to as encoding and the second one as decoding network. Taking a look at the example illustrated in Figure 18, the lemma of the German word “aufgetreten” is predicted depicting the encoding network in green and the decoding network in blue. At first the input word is regarded as a sequence being read character by character. Adding a new character as input value changes the internal state of the network which is illustrated by the different shades of green. Apart from processing characters the first network also receives the part of speech as an input value at the end of the sequence influencing the state as well. The final encoding state depicted in dark green is then passed on to the second recurrent neural network, which predicts the first character of the lemma, “a”. Based on the received output of the encoding network (dark green box), the internal state depicted in shades of blue, and on the previous prediction, the next character can be found, “u”. This procedure is repeated until the decoder network recognizes the end of the lemma.

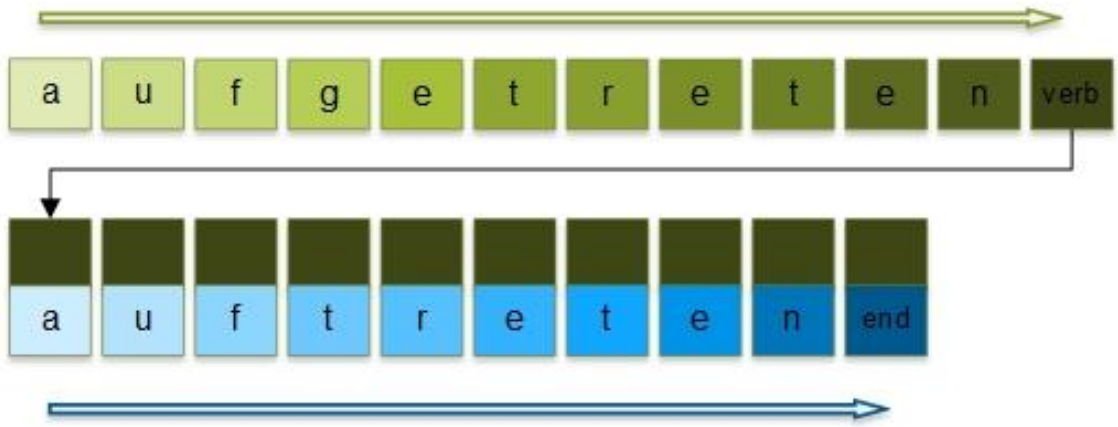

Figure 18: Lemmatization by sequential recurrent neural networks[116]

d.) Part-of-Speech Tagging

Part-of-Speech tagging (POS) is the task of labelling words by their word category and identification according to word context and properties. The simplest form of POS tagging is the classification in nouns, adjectives, verbs etc.[117] It is applied in various natural language processing models such as speech recognition or text parsing, drawing on a broad variety of methods including rule based approaches, Markov models, maximum entropy methods or machine learning techniques.[118]

A popular approach are hidden Markov models, assigning the POS tags based on probabilities. A hidden Markov model consists of finite states and probabilistic state

[116] cf. Cho/van Merrienboer/Gulcehre/Bahdanau/Bougares/Schwenk/Bengio, 2014, p.4
[117] cf. Nof, 2009, p.262
[118] cf. Anandarajan/Hill/Nolan, 2019, p.58

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

transitions, which means, starting from a given state there are various possible next states with a different probability of going to them. In POS tagging the state is represented as a pair (w, t) whereas w corresponds to a word and t to a part-of-speech tag such as noun, adjective or verb. On the word level there may be different POS- tags for each word. For example, “flies” could be a verb in the third person of fly but also a plural noun referring to the insect. Regarding an untagged sentence the sequence of words with their corresponding POS-tags represent states, however, the states are not observable due to the missing POS information or with other words the states are hidden. Hidden Markov models investigate all possible transactions between those states together with their transaction possibilities. These transaction probabilities can be derived from tagged language corpora by counting the occurring sequences of words and POS-tags. For example, the article “the” is frequently followed by a noun expressed by a high transaction probability. The procedure of finding the correct POS-tags of sentences should be clarified by the following example sentence. At first all possible POS tags for each word are listed.[119]

Table 3: Possible POS-tags of word examples

Word	Possible POS-tags
Flies	noun, verb
like	preposition, conjunction, adverb, noun, verb
a	article, noun, preposition
flower	noun, verb

Figure 20 illustrates all possible state transactions, which could occur in the given example sentences. This results in a tree diagram of which each branch corresponds to a state transaction with a given probability. Mathematically speaking the transaction can be expressed by the conditional probability of a possible future state given the current state. Multiplying the conditional branch probabilities along all possible paths delivers the joint probability for each path, revealing the most likely sequence of POS-tags.[120] In the given example the following sequences achieved the highest probabilities.

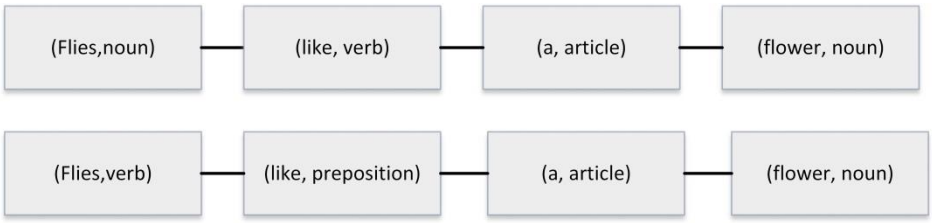

Figure 19: Word sequences with the highest joint probabilities

[119] cf. Nof, 2009, p.262
[120] cf. Ibid., p.262

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The introduced method relies on an assumption also referred to as Markov property. Being X_k a random variable expressing the state x_k, the sequence of *k* possible POS-tags corresponds to a sequence of *k* random variables. We assume that the probability of a future event only depends on the current state, which is also known as memoryless property. Thus, the conditional probability of a future state given all past and the current states equals the probability under the condition of the current state, which is expressed with the following Markov property equation:

$$\mathrm{P}(\mathrm{X_k} = \mathrm{x_k}|\mathrm{X_{k-1}} = \mathrm{x_{k-1}}, \dots, \mathrm{X_1} = \mathrm{x_1}) = \mathrm{P}(\mathrm{X_k} = \mathrm{x_k}|\mathrm{X_{k-1}} = \mathrm{x_{k-1}})$$

Equation 3: Markov probability equation

Moreover, the conditional probabilities are assumed to be invariant with respect to time, thus remain constant in the hidden Markov model.[121]

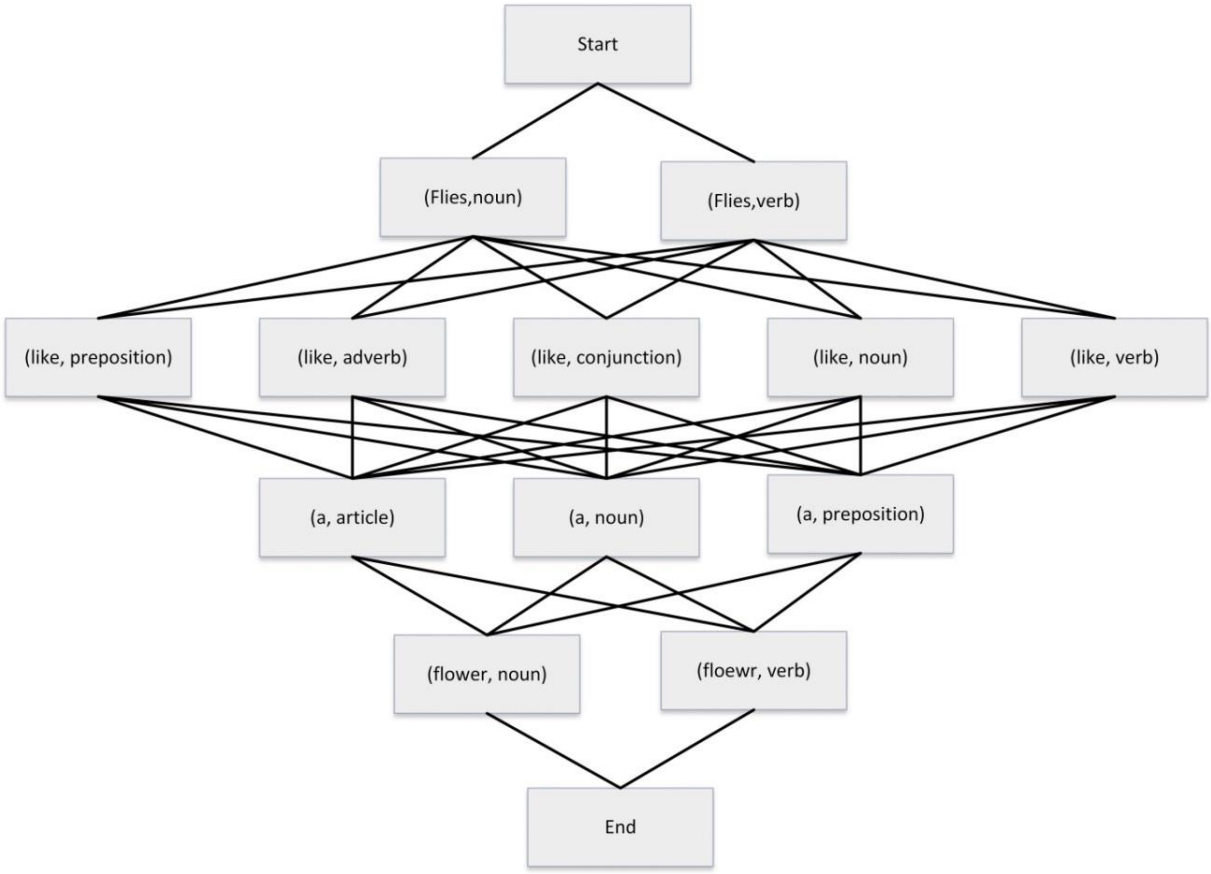

Figure 20: Visualization of all possible state transitions[122]

Another possibility for accomplishing POS-tagging are the previously introduced recurrent neural networks standing out by their ability to process sequential information. In Figure 21 an architecture for the task of POS tagging is illustrated being composed of an input -, hidden- and output layer. At first the tokens are converted to numerical representation, for example, by using one-hot-encoding

[121] cf. Martin/Reitz/Wehn, 2006, p.6
[122] vgl. Nof, 2009, p.262

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

vectors or embedding vectors (cf. Section 2.2.2) building the sequential input x_t. Since the sequence can be considered as a time series the variables carry the index t. The hidden layer and hidden states h_t represent contextual information learned from the input but also from previous hidden states. Then the probability distribution over tags of tokens can be derived from h_t using the softmax activation function (cf. Section 2.2.2). Finally, the output tag equals the tag which achieves the highest probability of the distribution.[123]

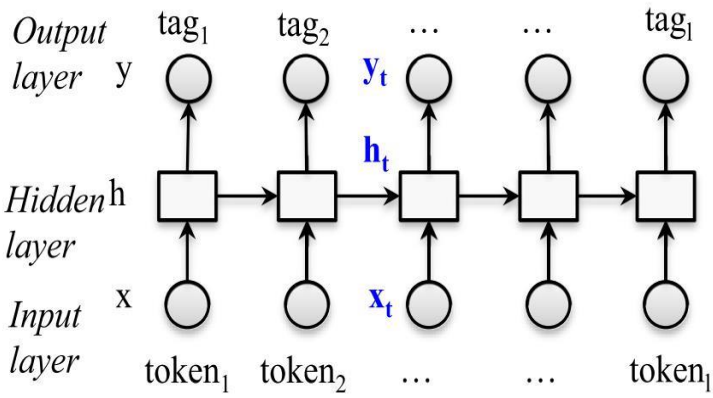

Figure 21: Recurrent neural network for POS tagging[124]

e.) Language Identification Annotator

Language identification is a task which belongs to the category of text classification assigning unstructured text to class labels. The implemented language identification annotator in the use-case utilizes Naïve Bayes classification and is thus focused on in this section. The approach is based on the Bayes' theorem shown in the formula below calculating the posterior probability by the product of prior and likelihood divided by the evidence.[125] Applied to the task of language detection the prior probability $P(L)$ describes the possibility of any text being in the chosen language. It is usually known and either depends on external factors or is assumed to be the same for all considered language categories. The evidence $P(C)$ is the probability of the characters of any random text being exactly as they are in the input string. It is constant for any fixed input. The probability of the input string under the condition of a given language is expressed by $P(C|L)$ and called likelihood. With the mentioned terms finally the required prior probability $P(L|C)$ can be determined; the probability of the language being L given the characters C.[126]

[123] cf. Nguyen/Nguyen/Tojo/Satoh/Shimazu, 2018, pp.175f
[124] Ibid., p.176
[125] cf. Tschirk, 2014, pp.12f
[126] cf. Winkelmolen/Mascardi, 2010, pp.15f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

$$P(L|C) = \frac{P(C|L) * P(L)}{P(C)}$$

Equation 4: Formula of Bayes' theorem

Statistical analysis for probability calculation prerequisites corpora in the languages of interest. The bigger and the more representative of the language the data is, the better the algorithm will perform. In order to derive the probabilities text is described by features. Features can be anything from n-gram frequencies to average word length, from vowel-consonant ratio to the number of two letter words. The most common and easy to use features are the frequencies of occurrence of words and n-grams. Having a text described by n features $F_1, F_2, \dots, F_n$ the probability of being in a certain language L is due to the Bayes' theorem equal to[127]

$$P(L|F_1, F_2, \dots, F_n) = \frac{P(F_1, F_2, \dots, F_n|L) * P(L)}{P(F_1, F_2, \dots, F_n)}$$

Equation 5: Bayes' theorem based on word features

The conditional probability $P(F_1, F_2, \dots, F_n|L)$ is calculated under the assumption of independent features $F_1, F_2, \dots, F_n$. Thus, the feature probabilities are regarded individually leading to:

$$P(F_1, F_2, \dots, F_n|L) = \prod_{i=1}^{n} P(F_i|L)$$

Equation 6: Likelihood under the assumption of independent features

This simplifies the model which is why it is also referred to as naive assumption giving the name to the classification method. Since the evidence $P(F_1, F_2, \dots, F_n)$ only depends on the features and not on the language classes, it can be seen as a constant K being indecisive for language classification. Thus, the final Bayes classifier assigns a class label based on the following equation:[128]

$$y = \underset{L}{\operatorname{argmax}}\ P(L) * \prod_{i=1}^{n} P(F_i|L)$$

Equation 7: Naive Bayes classification formula

f.) Word Distance Calculation

The word distance is a similarity measure expressing the minimum number of necessary insertions, deletions or substitutions for transforming one string to another and is also referred to as edit-distance or Levenshtein distance after Vladimir

[127] cf. Winkelmolen/Mascardi, 2010, p.16
[128] cf. Ibid., pp.17f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Levenshtein. It is applied for string matching and spelling correction and is due to its computational complexity appropriate for comparing short sequences. The algorithm for defining the minimal number of necessary edits being also implemented in the use case is given as follows:[129]

Being a, b two strings of length $|a| = n$, $|b| = m$ with i, j being their indices $0 \leq i \leq n$, $0 \leq j \leq m$ then a_i, b_j are the i-th character of a and the j-th character of b. Nested iterating over i, j and the following definition leads to a matrix $Lev_{a,b}$ of size $(n + 1)\ x\ (m + 1)$.

$$Lev_{a,b}(i,j) = \begin{cases} \max(i,j) & \min\,(i,j) = 0 \\ min \begin{cases} Lev_{a,b}(i-1,j) + 1 \\ Lev_{a,b}(i,j-1) + 1 \\ Lev_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & otherwise \end{cases}$$

Equation 8: Levenshtein distance matrix calculation

whereas the function $1_{(a_i \neq b_j)}$ equals 0 if $a_i = b_j$ and 1 if $a_i \neq b_j$. The Levenshtein distance of the words a, b finally is the last element in the matrix:[130]

$$Lev(a,b) = Lev_{a,b}(n,m)$$

Equation 9: Last element of Levenshtein matrix

The first statement in the definition $\max(i,j)$ if $\min\,(i,j) = 0$ simply assigns the first row and the first column of the matrix the indices sequences. The remaining three terms in the minimum function correspond to the actions of deletion, insertion and substitution each adding a cost of 1 to the left-, upper- or left upper element of the matrix and storing the minimum at $Lev_{a,b}(i,j)$.

- $Lev_{a,b}(i-1,j) + 1$: Deletion, cost added to left element
- $Lev_{a,b}(i,j-1) + 1$: Insertion, cost added to upper element
- $Lev_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)}$: Substitution: cost added to upper left element unless characters a_i and b_j are equal.

The computational complexity of that algorithm is given by the product of string length $\theta(m * n)$ growing rapidly with the number of characters. Thus, calculating the edit distance for long sequences can cause long run time and should be avoided. On the other hand Levenshtein distance stands out by its ability to compare strings of different length, which is not possible with simpler and less runtime consuming approaches such as the Hamming distance.[131] In the application field of spelling correction it is purposeful to compare the number of necessary edits to the maximal

[129] cf. Janowski/Mohanty, 2010, p.259
[130] cf. URL: https://web.stanford.edu/class/cs124/lec/med, 11.09.2019
[131] cf. Heeringa, 2004, pp.122f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

number of characters of the considered strings. This is referred to as genetic distance being represented by a value between zero and one.[132]

The following table shows the whole Levenshtein matrix investigating the edit distance between “Honda” and “Hyundai”. The blue background color marks the matrix indices represented in the first term of the definition; the red cell is the final outcome of Levenshtein distance.

Table 4: Example of Levenshtein matrix[133]

		H	Y	U	N	D	A	I
	0	1	2	3	4	5	6	7
H	1	0	1	2	3	4	5	6
O	2	1	1	2	3	4	5	6
N	3	2	2	2	2	3	4	5
D	4	3	3	3	3	2	3	4
A	5	4	4	4	4	3	2	3

2.2.2 Word Embedding and Representation

Word embedding describes a group of methods in text-mining aiming at converting words to a vector of real numbers providing a compatible format for analysis.[134] The numerical mapping should express the semantic and contextual relationship of words building the foundation for computational text comprehension.[135] There are different approaches for creating appropriate word representations including word frequency-based methods as well as mappings generated by neural networks. This work focusses on the two most common representatives of that categories referred to as Bag of words and Word2Vec models.

a.) Bag of Words

Bag of words is a concept which relies on the term composition of documents represented in a Term-Document Matrix. The definition of a document should be not confused with the general linguistics meaning of the term and refers to a more general meaning of text segments including also paragraphs or single sentences.[136]

The term document matrix is a two-dimensional array containing the term composition of each document. The rows represent a collection of different tokens or n-grams of tokens occurring in any of the documents, the columns correspond to the defined documents. The entries $a_{ij}$of the term document Matrix $A_{ij}$are then the frequencies of term i in document j. Notably, the term-document representation does

[132] cf. Janowski/Mohanty, 2010, p.259
[133] cf. URL: https://dzone.com/articles/the-levenshtein-algorithm-1, 11.09.2019
[134] cf. Furbach/Krämer/Schon, 2019, p.256
[135] cf. Ibid., p.256
[136] cf. Anandarajan/Hill/Nolan, 2019, p.60

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

not consider the order of words in the documents being a basic assumption of Bag of Words concept.[137]

For illustration, the following table of example sentences in un- and preprocessed form are numerically expressed by the term document matrix whereas each sentence represents an individual document. In

Table 6 the columns provide the vectoral Bag of Words representation of each document formed by the frequencies of tokens listed in alphabetically order.

Table 5: Un- and pre-processed example sentences for BoW representation[138]

Documents		
Number	**Text**	**Pre-processed text**
1	My favorite dig is fluffy and tan	[favorite] [dog] [fluffy] [tan]
2	The dog is brown and cat is brown	[dog] [brown] [cat] [brown]
3	My favorite hat is brown and coat is pink	[favorite] [hat] [brown] [coat] [pink]
4	My dog has a hat and leash	[dog] [hat] [leash]
5	He has a fluffy coat and brown coats	[fluffy] [coat] [brown] [coat]
6	The dog is brown and fluffy and has a brown coat	[dog] [brown] [fluffy] [brown] [coat]
7	My dog is white with brown spots	[dog] [white] [brown] [spot]
8	The white dog has a pink coat and the brown dog is fluffy	[white] [dog] [pink] [coat] [brown] [dog] [fluffy]
9	The three fluffy dogs and two brown hats are my favorites	[fluffy] [dog] [brown] [hat] [favorite]
10	My fluffy dog has a white coat and hat	[fluffy] [dog] [white] [coat] [hat]

Table 6: Example of Term-Document matrix[139]

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
brown	0	2	1	0	1	2	1	1	1	0
cat	0	1	0	0	0	0	0	0	0	0
coat	0	0	1	0	2	1	0	1	0	1
dog	1	1	0	1	0	1	1	2	1	1
favorite	1	0	1	0	0	0	0	0	1	0
fluffy	1	0	0	0	1	1	0	1	1	1
hat	0	0	1	1	0	0	0	0	1	1
leash	0	0	1	1	0	0	0	0	1	1
pink	0	0	1	0	0	0	0	1	0	0
spot	0	0	0	0	0	0	1	0	0	0
tan	1	0	0	0	0	0	0	0	0	0
white	0	0	0	0	0	0	1	1	0	1

[137] cf. Anandarajan/Hill/Nolan, 2019, p.63
[138] cf. Ibid., p.64
[139] cf. Ibid., p.64

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The values of the matrix are term frequencies expressing the importance of terms in the document which means the more often a word occurs in a document the more important it is to the content of that document. Referring again to our example, by using the term frequency the word brown in document 2 is considered to be twice as important as in document 3. Thus, the variability of frequencies causes strong differences and abrupt changes with respect to term importance which can be compensated by alternative weighting approaches.[140]

Being $f_{t,d}$ the frequency of the term t in document d and $tf(t,d)$ the local adjusted term frequency, the logarithmic scaling for reducing the effect of large differences is given as follows:

$$tf(t,d) = \begin{cases} \ln(f_{t,d}+1) & f_{t,d} > 0 \\ 0 & otherwise \end{cases}$$

Equation 10: Logarithmic scaled term frequency

Other alternative weighting possibilities are binary frequencies, measuring whether a term occurs in a document or not, without distinguishing between their different numbers of appearance.

$$tf(t,d) = \begin{cases} 1 & f_{t,d} > 0 \\ 0 & otherwise \end{cases}$$ [141]

Equation 11: Binary frequency-based weighting

Dividing the frequency by the number of words n of the underlying document is another approach for adjusting the word frequency difference taking into account the document length.

$$tf(t,d) = \frac{f_{t,d}}{n}$$ [142]

Equation 12: Local weighting adjustment based on document length

So far, the presented frequency measures focused on the on the word occurrence within a regarded document, referred to as local weighting. Respectively the importance of a term regarding the whole document collection is called global weighting.[143] A commonly applied global approach is the inverse document frequency investigating the number of documents in which a specific term occurs. It expresses the global word importance by alleviating the weight of terms being present in many documents scaled by the logarithm to the base of 10. N is the total

[140] cf. Anandarajan/Hill/Nolan, 2019, pp.64f
[141] cf. Ibid., p.66
[142] cf. Manning/Raghavan/Schütze, 2008, p.117
[143] cf. Anandarajan/Hill/Nolan, 2019, p.67

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

amount of documents, the expression $|d \in D: t \in d|$ refers to the number of documents which the term t occurs in.

$$idf(t, D) = \log\left(\frac{N}{|d \in D: t \in d|}\right)^{144}$$

Equation 13: Global weighting adjustment

Combining the principles of local and global weighting by multiplication leads to the term frequency inverse document frequency scheme (tf-idf). It is high if terms occur many times in few documents and low if a term almost appears in every document. Intuitively, words such as articles being strongly represented across all documents are less important than terms which occur frequently in just one or few documents.[145]

$$tf_idf = tf * idf$$

Equation 14: Term frequency - inverse document frequency weighting

Applying this weighting scheme to the set of example sentences gives a term-document tf-idf matrix, showing the word importance to a document. Apparently, the terms spot, cat and tan which occur infrequently across the document collection but frequently in a small set of documents show a high tf-idf value. On the other side the term dog is weighted by a low value due to its high global frequency and thus small idf value.[146]

Table 7: Example of tf-idf weighting

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
brown	0	3.03	1.51	0	1.51	3.03	1.51	1.51	1.51	0
cat	0	4.32	0	0	0	0	0	0	0	0
coat	0	0	2	0	4	2	0	2	0	2
dog	1.32	1.32	0	1.32	0	1.32	1.32	2.64	1.32	1.32
favorite	2.74	0	2.74	0	0	0	0	0	2.74	0
fluffy	1.74	0	0	0	1.74	1.74	0	1.74	1.74	1.74
hat	0	0	2.32	2.32	0	0	0	0	2.32	2.32
leash	0	0	0	4.32	0	0	0	0	0	0
pink	0	0	3.32	0	0	0	0	3.32	0	0
spot	0	0	0	0	0	0	4.32	0	0	0
tan	4.32	0	0	0	0	0	0	0	0	0
white	0	0	0	0	0	0	2.74	2.74	0	2.74

b.) Word2Vec

Word2Vec is a set of methods for capturing linguistic contexts of words by using a combination of two-layer neural networks. Each word is converted to a vector in a

[144] cf. Manning/Raghavan/Schütze, 2008, p.118
[145] cf. Anandarajan/Hill/Nolan, 2019, p.71
[146] cf. Ibid., p.71

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

semantic space enabling mathematical operations of linear algebra. Synonyms are represented by similar vectors quantifying the meaning of a word. With Word2Vec texts are converted to a matrix of dimension amount of words by amount of features building the word vectors.[147]

There are two different probabilistic concepts Word2Vec can be based on: Continuous bag of words (CBOW) using context words for predicting target words and Skip-gram being the inversion, thus using target words for predicting context words. Since both methods are algorithmically similar, this work focusses only on CBOW for explanation.[148] For facilitating demonstration it is again referred to the set of example sentence from above focusing on the tokens [favorite] [hat] [brown] [coat] [pink] trying to predict the word [hat] from the surrounding context words.

Figure 22 shows the architecture of CBOW consisting of a neural network with an input, hidden and output layer. Beginning with the input layer the variables x are the C regarded context words each being represented by a unit vector of which size corresponds to the number of unique words V occurring in the text. The word [brown] would be represented by a vector of size 12 consisting of a one in the first row and zeros in the remaining components $[1, 0 \dots 0, 0]^T$. All the other context words are analogously converted. Each vector is multiplied by the same weight matrix W_{VxN} whereas N is the dimension of the desired word vector representation being equal to the number of neurons in the hidden layer. Since in the example there are four context words, multiplying by the weight matrix results in four vectors of dimension N which are averaged to a single vector and passed on to the hidden layer. Notably, in this step there's no activation function such as ReLu or sigmoid and the output of the hidden layer corresponds to the weighted and averaged inputs. The values of W_{VxN} and W'_{NxV} are set randomly at the beginning.[149]

The weight matrix W'_{NxV} then maps the obtained outputs of the hidden layer to the output layer. The goal is to create an output vector of dimension V representing a probability distribution which assigns each word in the textual data base a probability for being the target word based on the regarded context words. As a consequence the components of the output vector have to add up to one which is achieved by applying the obtained outputs of matrix multiplication with W'_{NxV}, also referred to as activation $\boldsymbol{u}'$,to the softmax function. The softmax function is an exponential function normalizing the components of a real vector to an interval between zero and one. A single component y_j of the output vector can be computed as follows:[150]

[147] cf. Alexandrov/Boukhanovsky/Chugunov/Kabanov/Koltsova, 2018, pp.314f
[148] cf. Rebala/Ravi/Churiwala, 2019, pp.123f
[149] cf. Pattanayak, 2017, pp.228f
[150] cf. Ibid., pp.228f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

$$y_j = \frac{e^{u'_j}}{\sum_{n=1}^{V} e^{u'_n}}$$

Equation 15: Softmax function

The presented steps so far relied on the randomized values of W_{VxN} and W'_{NxV}. Thus, the obtained output is compared to the desired output for building an error function and adjusting the weights by backpropagation as explained in the previous Section 2.2.1. In the mentioned example the desired output of the context words [favorite] [hat] [brown] [coat] [pink] is [hat] which corresponds to the probability distribution $[0,0,0,0,0,0,1,0,0,0,0,0]^T$. After the procedure has been applied to the whole textual training set the weight matrix W_{VxN} can map each word of the corpus to the word embedding of the chosen N dimensions.

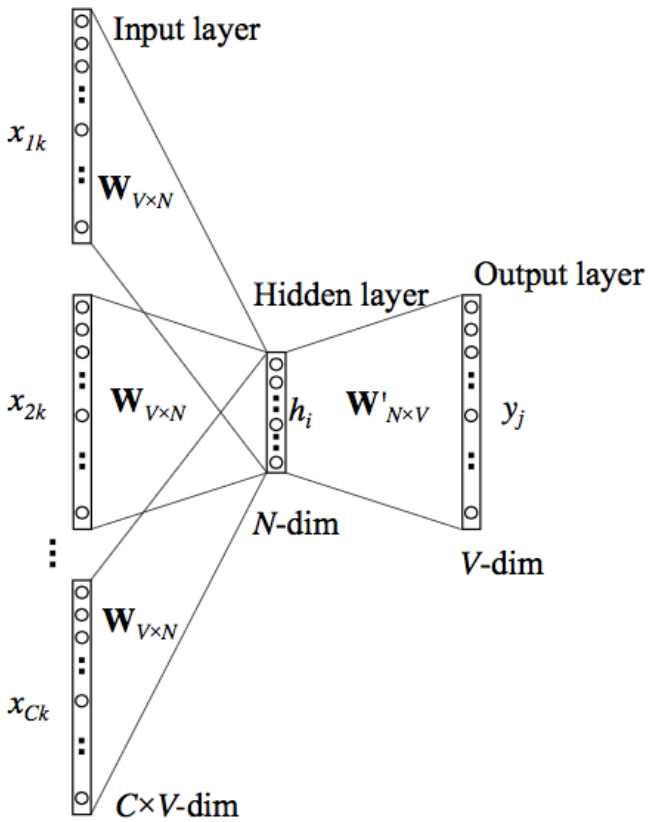

Figure 22: General architecture of Continuous Bag of Words[151]

After having accomplished the embedding of words, their semantic distances can be calculated. A common method for measuring the similarity between embedded words is the Cosine similarity based on the definition of the dot product. The cosine of the angle between two regarded vectors, as shown in Figure 23 equals their dot product divided by the product of their vector lengths.[152]

[151] cf. URL: http://www.stokastik.in, 16.09.2019
[152] cf. Vezzetti/Marcolin, 2015, p.47

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

$$similarity = \cos(\alpha) = \frac{\boldsymbol{a} * \boldsymbol{b}}{\|\boldsymbol{a}\| * \|\boldsymbol{b}\|}$$

Equation 16: Cosine similarity of two vectors

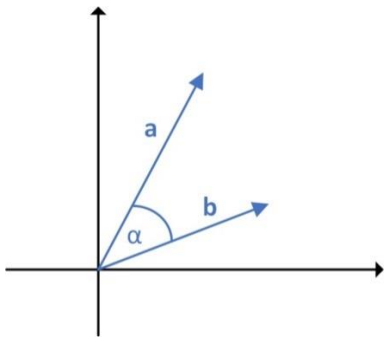

Figure 23: Cosine similarity measurement

2.2.3 Association Measuring

Association Measuring is an approach for imitating the mental ability of humans to build associations and was developed to improve knowledge extraction from text documents with multitopic structures. Conventional methods for keyword extraction provide a general content overview detecting a specific topic in text. Regarding large text corpora this results in numerous key words with multiple topics. Association Measuring is used to numerically express the association between those key words and serves as foundation for calculating their strength of attraction and for detecting multitopical structures.[153]

One main challenge of association measuring is the asymmetrical attribute of human association, which means the strength of association between two words depends on the starting point and direction. Thus, the following concept was derived from a large number of human word association case studies with special focus on the asymmetrical aspect. The derived equation is called CIMAWA standing for Concept for the Imitation of the Mental Ability of Word Association and is given by:[154]

$$CIMAWA_{ws}^{\zeta}(x(y)) = \frac{Cooc_{ws}(x,y)}{(frequency(y))^{\alpha}} + \zeta * \frac{Cooc_{ws}(x,y)}{(frequency(x))^{\alpha}}$$

Equation 17: CIMAWA word association concept

CIMAWA quantifies the association of word x with word y, whereas x represents the stimulus meaning x is the word which associated terms should be found for. The expression ws defines the window size for co-occurrence measuring counting the mutual occurrence of word x and y. The total word frequencies $frequency(x)$ and

[153] cf. Klahold/Uhr/Ansari/Fathi, 2014, p.40
[154] cf. Ibid., pp.40f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

$frequency(y)$ are exponentiated by the parameter alpha and have a high impact on CIMAWA especially in case of low frequencies. For overcoming that difficulty, a case differentiation is made based on the following boundaries:[155]

$$\frac{Cooc_{ws}(x,y)}{(frequency(y))^{\alpha}} = \begin{cases} \frac{Cooc_{ws}(x,y)}{(frequency(y))^{\alpha}}, & frequency(y) > \beta * Q \\ \frac{Cooc_{ws}(x,y)}{\gamma * Q}, & frequency(y) \leq \beta * Q \end{cases}$$

Equation 18: Case differentiation of CIMAWA

Q depends on the analyzed corpus and represents the total number of words. Based on empirical research the best results are achieved with $\alpha = 0.68$; $\beta = 5 * 10^{-6}$; $\gamma = 5 * 10^{-6}$. Analogously the case differentiation is made for the word frequency of x in the second part of the CIMAWA equation.[156]

As apparent in the formula, the derived concept of CIMAWA is a hybrid approach linearly combining the two association directions by a damping factor ζ. The dampingfactor controls the strength of asymmetry, which means $\zeta = 1$ results in completely symmetric behavior implying $CIMAWA_{ws}^{\zeta}(x(y)) = CIMAWA_{ws}^{\zeta}(y(x))$. On the other hand, $\zeta = 0$ turns CIMAWA into a completely asymmetric model. The mentioned overlap of association directions and the concept behind the CIMAWA equation are depicted in Figure 24. Comprehensive case studies of free association tests compared to CIMAWA based on large corpora have shown that a dampingfactor with an interval of [0.4-0.6] provides the best results. Thus, in the use-case the damping factor is set to 0.5 being a promising assumption.[157]

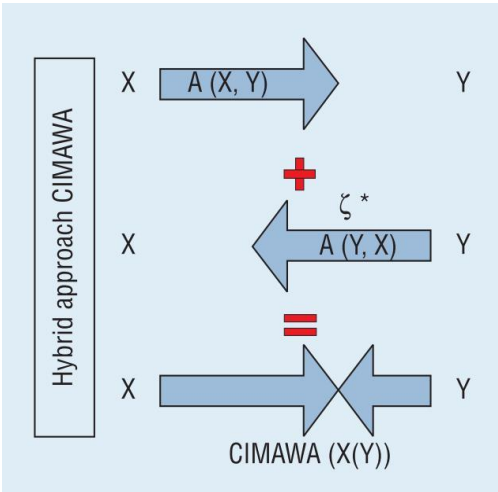

Figure 24: Visualization of CIMAWA concept[158]

[155] cf. Uhr, 2014, pp.40f
[156] cf. Ibid., pp.40f
[157] cf. Klahold/Uhr/Ansari/Fathi, 2014, p.41
[158] Ibid., p.41

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.

In order to fulfil the goal of detecting multitopical structures and building topic clusters, the CIMAWA model is extended by taking into account the importance of extracted key words. The extended model is referred to as Associative Gravity Force being inspired by the physical law of gravitation. The theory of gravitation postulated by Isaac Newton defines the attraction between each particle of matter as a force directly proportional to the product of their masses and inversely proportional to the square of the distance between them. Although Isaac Newton's law of gravity shows behaviors which can't be directly transferred to the proposed model, such as the symmetry of force effect being not consistent with the asymmetry of association, parallels can be found which are decisive for the naming. The mass of the particles corresponds to the importance of the key words expressed by the tf-idf scheme; the distance between them is the complement to the association calculated by CIMAWA. Besides the quantification of the key word importance, the tf-idf calculation serves as indicator for key word extraction. After the next step of CIMAWA calculation, the associative gravity force (AGF) of the extracted keywords can be determined by following equation:[159]

$$AGF\big(word_1(word_2)\big) = \frac{CIMAWA_{ws}^{\zeta}\big(word_1(word_2)\big) * kr(word_1)}{kr(word_2)}$$

Equation 19: Associative Gravity Force

The AGF represents the attraction between the key words and is based on the concept of CIMAWA extended by the word importance kr corresponding to the tf-idf values. Applying this method to a body of text documents reveals a network of keywords being attracted by asymmetric forces. Figure 25 shows a possibility of visualizing the introduced concept. The diameter of the keywords circles is direct proportional to the importance measured by the tf-idf value. The thickness of arrows represents the attraction numerically evaluated by the AGF equation.

The proposed concept provides a sophisticated approach for dealing with large amounts of unstructured textual data being not directly accessible for computational processing. It enables the analysis of multitopic text by extracting key words and detecting their attraction. Depicting the results by the illustrated visualization method allows the clear communication of that hidden knowledge which could be valuable for decision support. Moreover, implementing appropriate cluster algorithms could achieve a reduction of complexity by aggregating keywords based on their AGF value.

[159] cf. Klahold/Uhr/Ansari/Fathi, 2014, p.43

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

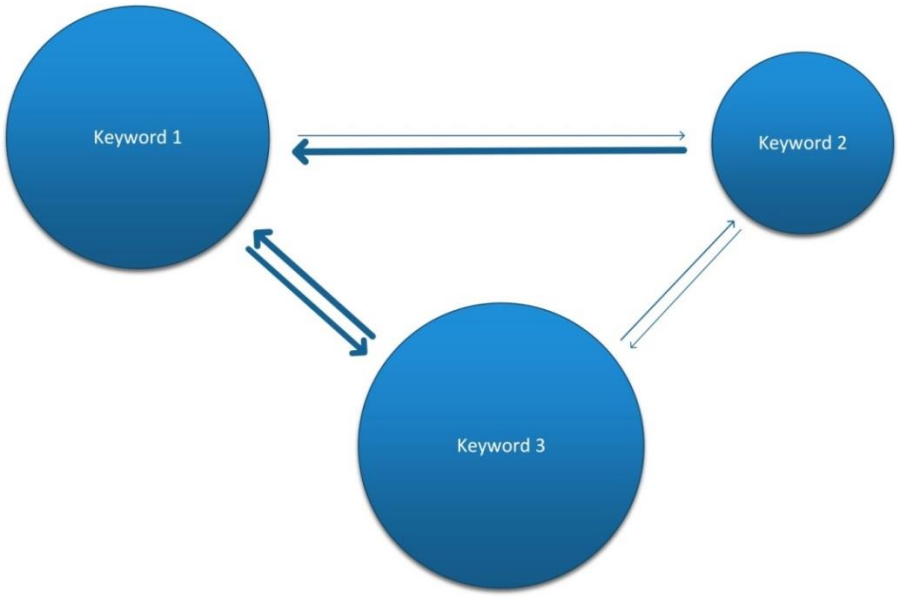

Figure 25: Visualization of Associative Gravity Force[160]

2.2.4 Opinion-Mining and Sentiment Detection

Sentiment analysis being also referred to as opinion mining is the computational evaluation of person's attitudes, opinions, appraisals and emotions towards topics, events, issues, individuals, events or entities. It is commonly used for extracting consumer opinions of product reviews providing valuable feedback for businesses across various sectors. With the rapid growth of social media including for example blogs, forum discussions and reviews, public opinions towards products and services gained importance for decision making. These huge amounts of available texts exceed the human capability of analysis and reinforced the application of computational opinion mining. Moreover, automated mining avoids human biases of subjectivity such as people's affinity to attitudes which are consistent with the own opinion. Thus, computational sentiment analysis is needed for imitating the human capability of capturing opinions providing an objective analysis system whilst being able to process big amounts of data.[161]

Before starting with the investigation of different approaches for sentiment analysis, it is purposeful to provide a general definition and clarify the properties characterizing an opinion. An opinion can be described as a vector $(e_i, a_{ij}, oo_{ijkl}, h_k, t_l)$ where $e_i$is the entity representing the target of an opinion. An entity can be a product, service, person, event, organization, or topic being described by a set of attributes a_{ij}. An entity could be for example a mobile phone, the attributes the weight and the seize characterizing it. oo_{ijkl} is the opinion orientation with respect to attribute a_{ij} of entity e_i. It is classified in positive, neutral or negative categories or continuously quantified

[160] cf. Klahold/Uhr/Ansari/Fathi, 2014, p.42
[161] cf. Aggarwal/Zhai, 2012, p.415

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

by intensity levels. The holder of an opinion is expressed by h_k, which in reviews and blogs usually also corresponds to the author of the text. Finally, t_l is the time the opinion is expressed by h_k. The indices show the internal dependencies of the introduced opinion components revealing the complexity of architecture in sentiments. The described concept provides a comprehensive definition of opinions of which components are important for meaningful sentiment mining. However, valuableness of information depends on the respective application which means the introduced vector doesn't necessarily have to be extracted in full extend and attributes should be selected according to the individual requirements.[162] A common assumption for simplifying the analysis of product and service reviews is the focus on a single entity and a single opinion holder and the investigation with respect to the aspect general. This implies that e, h and t are known or irrelevant and a is set to general.[163]

a.) Sentiment Analysis based on Supervised Learning

Besides the application field of pre-processing supervised machine learning also plays an important role in the environment of sentiment analysis. Supervised machine learning is based on a labeled training set and searches for patterns and correlations in the underlying data. Sentiment Analysis can be seen as a classification problem, allocating opinionated texts to the categories negative, neutral and positive. This task can be accomplished by almost any existing supervised learning methods including the introduced neural networks and naïve Bayesian classifier. An essential component for achieving successful machine learning is the selection of input features, also known as feature engineering. For detecting the sentiment of texts, the following input features are supposed to enable an effective classification towards the sentiment of texts:

- Terms and their frequency: The tf-idf weighting concept used for word embedding and measuring the importance can be applied to single words but also to n-grams.
- Part-of-speech: According to researches adjectives carry valuable information for sentiment classification and are treated as special indicators.
- Opinion words and phrases: Opinion words are words which are commonly used for communicating sentiments such as "wonderful", "good" and "amazing" but also "bad", "poor" and "terrible". Although adjectives represent the group with the most opinion words, there are also nouns such as "rubbish" and "junk" or verbs such as "hate" and "like" which indicate sentiments. On the level of phrases opinions often are expressed implicitly for example by idioms

[162] cf. Aggarwal/Zhai, 2012, p.418
[163] cf. Ibid., p.422

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

such as “cost someone an arm and a leg” which should be recognized and considered in the evaluation too.

- Negations: Negations imply a change of orientation in the sentiment and are thus indispensable in model creation and feature selection. However, detecting opinion shifters might appear more challenging than expected. For example, the word “not” in the sentence structure “not only… but also” does not change the orientation.[164] One approach for dealing with that problem could be the integration of negations into a semantic representation of already known relations instead of isolated consideration.[165]
- Syntactic dependency: The generation of word dependency trees provides information about phrase structures and their relations and can serve as foundation for input values of supervised machine learning. Dependency trees rely on probabilistic models drawing statistical occurrence measurements of texts or on deterministic bottom-up approaches investigating the dependency upon lexical information.[166]

The creation of labeled training data set being a prerequisite for supervised machine learning can be time consuming and come along with big manual effort. To cope with that challenge opinion words can be used to reduce that effort and detect appropriate labels with computational support. Based on the selected features and the applied word embedding, the machine learning algorithm learns from the labeled training data set and derives a function which finally enables the classification of new unlabeled examples.

To mention a recent example, in Figure 26 a novel approach for achieving sentiment analysis by convolutional neural networks is depicted. The illustration provides a simplified overview of the model and the following explanation focusses on the main steps necessary for fundamental understanding.

[164] cf. Aggarwal/Zhai, 2012, pp.423f
[165] cf. Blanco/Dan, 2011, p.230
[166] cf. Yamada/Matsumoto, 2003, p.1

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

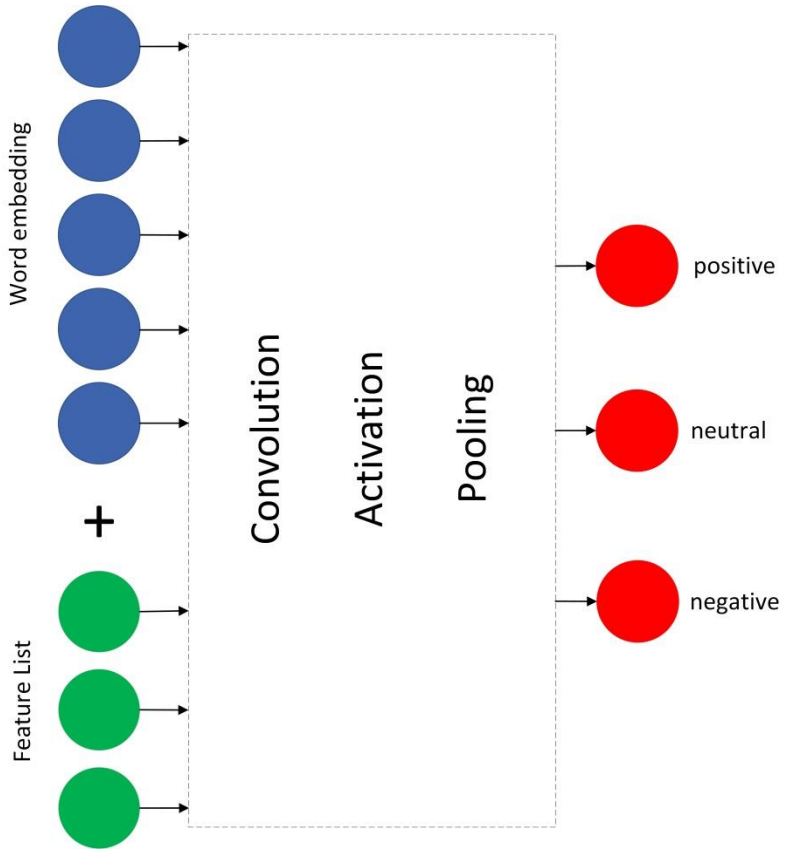

Figure 26: Model of a convolutional neural network for sentiment analysis[167]

It starts with an input vector consisting of the numerical word embedding representation, such as word2vec appended to a feature vector providing additional information about author and text. Convolutional neural networks are different from classical feed forward networks due to their typical architecture consisting of convolutional-, pooling- and activation layer being apart from sentiment analysis commonly applied to machine vision tasks.[168] The neurons of the convolutional layer are only connected to a limited area of neurons of the previous layer reducing the dimensionality, which means in this layer only pixels or input values within a certain area affect each other. This architecture acts as filter being systematically applied by sliding through all input values for extracting patterns such as edges of images. This procedure is systematically repeated by applying different filters of which each one searches for specific features resulting in a so-called feature map. The input values thus become a stack of convoluted values characterized by the applied filter of each layer as illustrated in Figure 27.[169]

In the activation layer the values are transformed by the activation function, which is commonly the rectifier (ReLu) turning negative values to zero and leaving positive values unchanged. That models the non-linearity of the features which could be

[167] cf. Alharbi/Doncker, 2019, p.54
[168] cf. Ibid., pp.53f
[169] cf. Aghdam/Heravi, 2017, pp.90f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

borders in pictures.[170] Related to the depicted model the activation step is not explicitly mentioned because it can be also seen as final part of the convolutional layer.

The pooling layer eliminates dispensable information by regarding the filtered values through a predefined window and combining them to one value by a function for shrinking the complexity and dimensionality. The function can be for example the maximum or the average of the values occurring within each window.[171]

Since there is still a stack of activated and pooled feature maps, the values are gathered in one layer also referred to as fully connected layer. Each of the values contributes to the decision whether the classification is positive, neutral or negative and therefore each of them is connected to the each node of the output layer.[172] Applying backpropagation finally achieves defining the weights, representing the numerical expression of the different filters and the extent of contribution of the values in the fully connected layers with respect to the final classification.

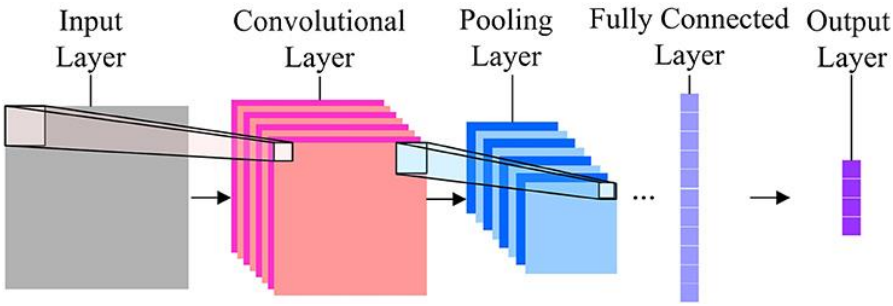

Figure 27: General model of a convolutional neural network[173]

b.) Sentiment Analysis with Opinion Lexica

As already mentioned in preceding sections, opinion words are essential in sentiment analysis and build the foundation for many applications. There are positive opinion words, such as beautiful, wonderful or amazing, and negative opinion words such as poor or terrible. Moreover, idioms can also implicitly express a positive or negative sentiment e.g. "cost someone an arm and a leg" obviously belonging to the negative category. For generating such opinion word lists there are three main approaches, manual-, dictionary and corpus-based approach. Creating a dictionary manually is very time consuming but accurate in classification with respect to the application

[170] cf. Aghdam/Heravi, 2017, pp.142f
[171] cf. Ibid., pp.95f
[172] cf. Ibid., pp.85f
[173] cf. Krizhevsky/Sutskever/Hinton, 2012

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

environment. Thus, a combination with automated approaches is purposeful improving time efficiency and maintaining accuracy.[174]

The dictionary-based approaches use bootstrapping, starting with a small group of manually selected opinion words which are well-known with respect to their orientation. This set of words is grown by using thesauri providing synonyms and antonyms for each word. The newly found words are added to the corresponding word list before starting the next iteration of search. This procedure is repeated until no more new words are found. Finally, the obtained lists can be manually inspected to correct possible errors. Moreover, additional information could be added e.g. glosses for improving the extent and accuracy of word lists. This approach has drawbacks if it is applied to various fields due to existing domain specific orientations. For example, the word quiet could imply a positive sentiment in relation with a car, but express negative properties in combination with a speaker of a phone. On the other hand, it's a simple and effective approach within a special domain, since it can be easily adapted and controlled by the manual selection of characteristic seed words.

The corpus-based approach also relies on a seed list of words but extends them by considering linguistic constraints and sentiment consistency in a large corpus. A linguistic constraint for example enables the classification of adjectives being linked by the conjunction "and". Taking a look at the sentence "This car is beautiful and spacious", if the word "beautiful" is a positive opinion word it can be implied that "spacious" belongs to the positive category as well. The reason for that is the assumption that only adjectives with the same orientation are combined by the conjunction "and". Unequal orientations linked by the conjunction "and" would in the most cases appear strange as the following sentence should illustrate: "The car is beautiful and difficult to drive". Of course, if the conjunction is replaced by "but" the sentence becomes acceptable leading to a constraint of inverted orientation. This rule of sentiment consistency can be derived for various conjunctions and connectors such as "or", "either-or" and "neither-nor". Based on the chosen seed words and the implemented constraints the orientation between adjectives can be learned from a large corpus. Applying clustering finally separates them in a positive and a negative word list.

After the generation of positive and negative opinion word lists, the created dictionaries can be used for sentiment analysis of texts. At first it has to be ensured that the text document is pre-processed ensuring unbiased word comparisons. Especially the methods of tokenizing, lemmatizing, lowercasing and spelling correction are advisable in order to harmonize multifarious words. Then the prepared tokens are compared with the created opinion word lists identifying the number

[174] cf. Aggarwal/Zhai, 2012, p.429

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

positive and negative words. The resulting distribution finally serves as measure for the sentiment of analyzed document. Tokens which are neither positive nor negative can be assigned to the neutral category or be removed from the underlying set for proportion calculation. The following figure shows an analysis of 3122 pre-processed tokens of which 1080 are positive and 1473 negative. In this analysis neutral tokens are not explicitly mentioned.

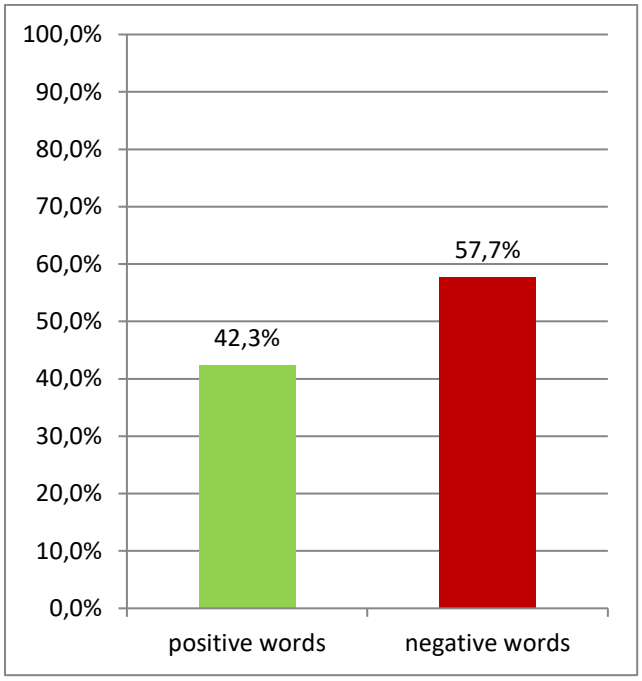

Figure 28: Distribution of negative and positive opinion words

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.

3 Practical Application of Text Mining: State-of-the-Art

3.1 Text-Mining Approaches applied in Various Business Sectors: A Brief Overview

The amount of unstructured textual data in business environments is growing rapidly being found in forms such as e-mails, logs, electronic health records, customer reviews and comments. While the amount of data increases, the mining, understanding and summarizing of them for supporting decision making and gaining competitive advantages stays a challenging task. In this section, text-mining applications are investigated from a general perspective covering different business sectors such as customer intelligence, health care and finance. Their methods and outcomes are briefly elaborated and finally summarized in Table 9.

Chakraborty and Pagolu propose a set of different methods with a focus on customer intelligence for effectively drawing benefits from this huge amount of available textual data.[175] First of all, text analysis can be utilized for improving the predictive accuracy of numerical data models relying on methods for prediction such as multiple regression or neural networks. The authors suggest Singular Value Decomposition for converting text to numerical representation for reducing the dimensionality and obtaining additional input data based on the textual sources. Especially, customer reviews are a promising source containing customers' experiences and opinions. Under the assumption that customer's experiences are consistent with the decision to continue with the business or to churn, the converted text combined with other numerical attributes, such as sales figures, serve as input for the neural network for predicting customer behavior. It has shown that the consolidation of numerical and textual data improves the predictive power in churn modeling and supports decision making in customer relationship management. The study is based on a regression model used for variable selection followed by an artificial neural network model (ANN) to predict customer's opinion. As shown in Table 8, involving textual data as additional input achieved a reduction of misclassification as well as an increase of sensitivity. [176]

Table 8: Misclassification and sensitivity rates of customer churn prediction[177]

Model	Misclassification	Sensitivity
Reg+ANN (Num+Txt)	18.75%	82.86%
Reg+ANN (Num)	26.56%	60.00%

[175] cf. Chakraborty/Pagolu, 2014, p.1
[176] cf. Ibid., p.2
[177] cf. Ibid., p.3

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

In the approach of Thiel et al. customer intelligence is created from social media data by combining methods of Network Analytics for quantifying the importance of nodes in a network and text-mining. In the application field of social media data network analysis is utilized for extracting relationships between individuals based on their communication. People who post articles with lots of response and who initiate discussions are identified as leaders. They are in the center of attention and play an import role in opinion making. On the other hand, followers comment a lot but don't receive many comments on their posts. That behavior of individuals can be measured and quantified by authority and hub weights, expressing their leader and follower score. Additionally, the sentiment of comments is analyzed using dictionaries of opinion words. Based on the positive, neutral and negative classification the sentiment distribution is used to determine the attitude of users. Applying cluster algorithms reveals finally similar groups of users with respect to sentiment and leadership which provides valuable information for company's marketing strategies and personalized advertising.[178]

Another possibility for knowledge extraction from text is the Root Cause Analysis. It aims at finding reasons for occurring failures and is for example applied for detecting inferences between potentially faulty products and reported health issues. In the regarded approach emergency room data provided by an electronic surveillance system is analyzed with a special focus on injuries associated with the keywords of “falling”, “sports” and “swallowing”. After applying the pre-processing methods of stemming and synonym recognition, the concept of tf-idf identifies keywords and drops unimportant terms. By applying association measuring the strength of inference between the focused injuries and extracted key words is measured quantifying the connection to potential root causes and products. The obtained dependencies can then be visualized in a linking diagram showing the key terms associated with for example the word “fall”. The association measuring is based on the conditional probability expressing the probability that term B exists given the term A already was found in the document. Although the strength of association is a good indicator for word inferences, high association strength does not necessarily imply causal relationship. As a consequence, visualized results have to be evaluated and interpreted by domain experts possessing the required background knowledge. Apart from detecting products associated with certain injuries, the approach exposes relations to actions and attempts to understand different types of injuries.[179]

The broad field of application of text-mining becomes apparent in the following approach of Williams and Gong. They developed a model for predicting the construction cost overruns using text-mining, numerical data and ensemble classifiers. The investigated textual data are project summaries describing the

[178] cf. Thiel/Kötter/Berthold/Silipo/Winters, 2012, p.3
[179] cf. Chakraborty/Pagolu, 2014, p.6

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

construction object and the project line. Important numerical parameters are the project's magnitude and the number of bidders in the tendering. After defining the three classes of cost overrun, project texts are pre-processed and converted to numerical representation using the BOW concept. Additionally, applying SVD reduces the dimensionality of the sparse matrix and increases computational efficiency. Numerical data and converted textual data form then the input for the classification being conducted by different techniques such as Radial Basis Functions being especially useful for problems with many variables. It uses the k-means algorithm for clustering, whereas symmetric multivariate Gaussians are fit to the data of each cluster building the basis-functions. The approach combining textual and numerical data for cost overrun showed high precision and recall and definitely improved predicting overruns. Construction projects with under runs showed less accuracy at first but could be improved by using stacking prediction which learns from the outputs of various classification methods and serves as a second classification instance.[180]

The mining of textual documents is also applied in the financial sector. Text-mining approaches are used for stock market prediction by analyzing the impact of news on the stock market behavior. The challenging task of stock market prediction can be modelled as a machine learning problem, inferring correlations from pre-classified news and market information such as market prices. Nikfarjam, Emadzadeh and Muthaiyah structure the task in three different components: News labelling, Classifier input generation and Classification. The available textual source of news articles is labelled based on the percentage change of market prices 15 minutes after news release. News articles co-occurring with an increase of at least three percent is assigned the label “buy”, whereas articles with a decrease of more than three percent are labelled as “short”. Classifier input generation includes the task of key word selection for representing the whole document. There are manual methods requiring financial experts for defining representative terms for each category of labelled articles. As computational alternative the BOW concept with tf-idf weighting is mentioned being applied to pre-processed tokens. Classification is accomplished by Support Vector Machine as binary classifier dividing the training data points by a hyperplane. The parameters and positioning of the hyperplane is derived by maximizing the distance to the nearest data points of each category. Depending on the number of categories and the chosen features for article representation the SVM approach shows directional accuracies between 45% and 70 %. It is pointed out that improvements could be achieved by considering technical indicators for market analysis instead of labelling news only based on pure market prices.[181]

[180] cf. Williams/Gong, 2014, p.23
[181] cf. Nikfarjam/Emadzadeh/Muthaiyah/Mahadevan, 2010, pp.256f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

In the sector of health care text-mining can be implemented as an additional surveillance approach for identifying postoperative complications by analyzing electronic medical records. In an interdisciplinary approach, conducted by the American Medical Association, electronic records of 2974 patients of six VHA medical centers are analyzed. They use an available ontology of medical nomenclature providing more than 300,000 hierarchical concepts for indexing the free text records. Together with rules generated by clinical teams and existing patient safety indicators, postoperative complications can be identified using string searches based on six predefined complication categories. Complication categories for instance are sepsis or pneumonia. Rules include the occurrence of typical key words such as “immunocompromised” but also word number combinations such as white blood cell count >12 000/µL. The generated NLP model acts as a classifier achieving satisfying accuracies across all postoperative complication groups with proportions between 59%-91%. Compared to existing approaches relying on the discharge code, the implementation of text-mining achieves a significant increase in sensitivity with a small reduction in specificity. [182]

3.2 Applications of Text-Mining in Production Systems

Although text-mining applications in production systems are still scarce, there are approaches for coping with the challenge of the increasing diversity of variants caused by the rising demand of customized products. Jiao et al. investigate similarities of products and their operation routings for exploiting common processes and reusing existing facilities. The goal is to reduce costs by enforcing operation repetitions and relying on standardized and proven process elements. The proposed model consists of three stages: Similarity measuring comparing the operation descriptions as well as tree structures of operation sequences, routing clustering building groups of similar routings, and routing unifications deriving master routings and a common process structure of each cluster. In Figure 29 two different operation routings are illustrated consisting of operations involving primitive components depicted in blue and compound components depicted in yellow. Each node additionally contains operations characteristics which are represented in textual descriptions. Apart from their tree structure and sequence of operations, the content of each node has an impact on their degree of similarity. Thus, the similarities of different product routing’s operations have to be extracted by suitable text-mining methods. After preparing text files by various pre-processing steps, primitive components and their corresponding descriptive attributes, such as weight and shape information occurring in the operating description, are extracted as key words. Frequency measuring then reveals how often an attribute is used to characterize individual primitive components indicating the importance of an attribute expressed

[182] cf. Murff/Fitz-Henry/Matheny/Gentry, 2011, pp.848f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

by the proportional weight. The similarity degree can then be determined by converting words to domain specific value representation and calculating the weighted attribute differences. Based on a bottom up approach of bipartite matching the obtained similarity degree is used to determine commonalities also between compound components. Together with the non-textual similarity measures and the remaining clustering and unification stages, so called generic routings can be identified representing common master processes across a product family.[183]

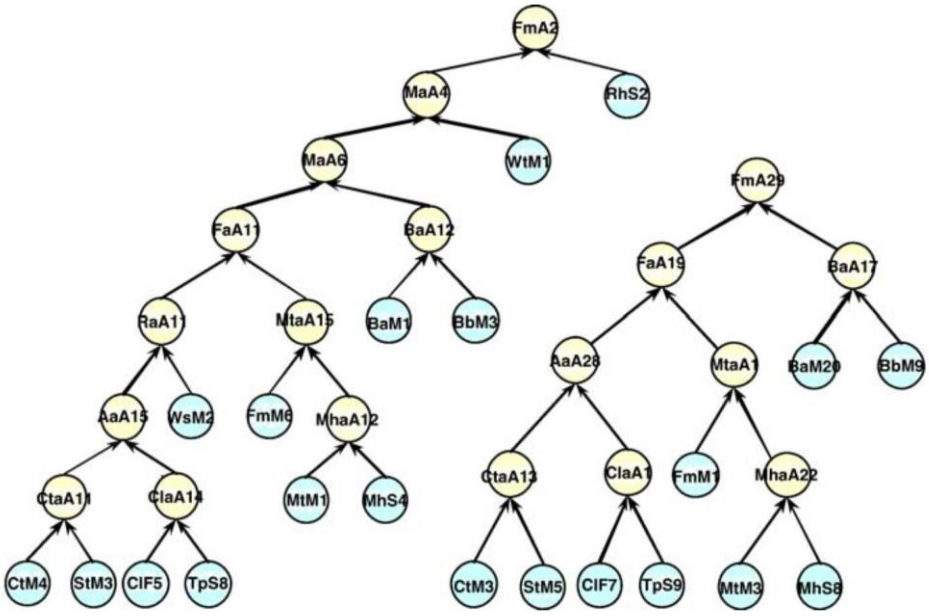

Figure 29: Visualization of two different operation routings[184]

The approach of Liu and Ke is dedicated to general field of problem solving in production by combining knowledge discovery from text and case-based reasoning. It is pointed out that conventional techniques of case-based reasoning provide solutions to similar problems without considering the actual information requirements of workers. The proposed approach, thus, aims at extending case-based reasoning by introducing a knowledge support system for workers enabling root cause detection of occurring problems. The model procedure is illustrated in Figure 30. The underlying information sources are historical problem-solving logs containing documents accessed by workers in specific situations. In the stage of “data preprocessing for knowledge discovery” the tf-idf concept is applied for transforming the documents to weighted term vectors which are used by the following “profile discovery” module to generate situation/action profiles. Similarity measuring of the current case or situation with existing situation/action pairs represented in weighted term vectors builds then the foundation for document retrieval with case-based reasoning. Additionally, in the module of “knowledge pattern discovery”, association rule mining derives inferences between conducted actions and specific situations

[183] cf. Jiao/Zhang/Pokharel/He, 2007, p.866
[184] Ibid., p.870

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

serving as decision support and helping workers to take appropriate actions. Association rule mining utilizes the statistical measures of support and confidence for describing implications from one item to another. Support expresses the proportion of transactions which contain both items whereas confidence is defined as the fraction of transactions containing the first item that also contain the second one. The results have finally shown that the application of text-mining enriches the existing case-based reasoning approaches by not only identifying similar situations but also discovering rules for supporting decision making.[185]

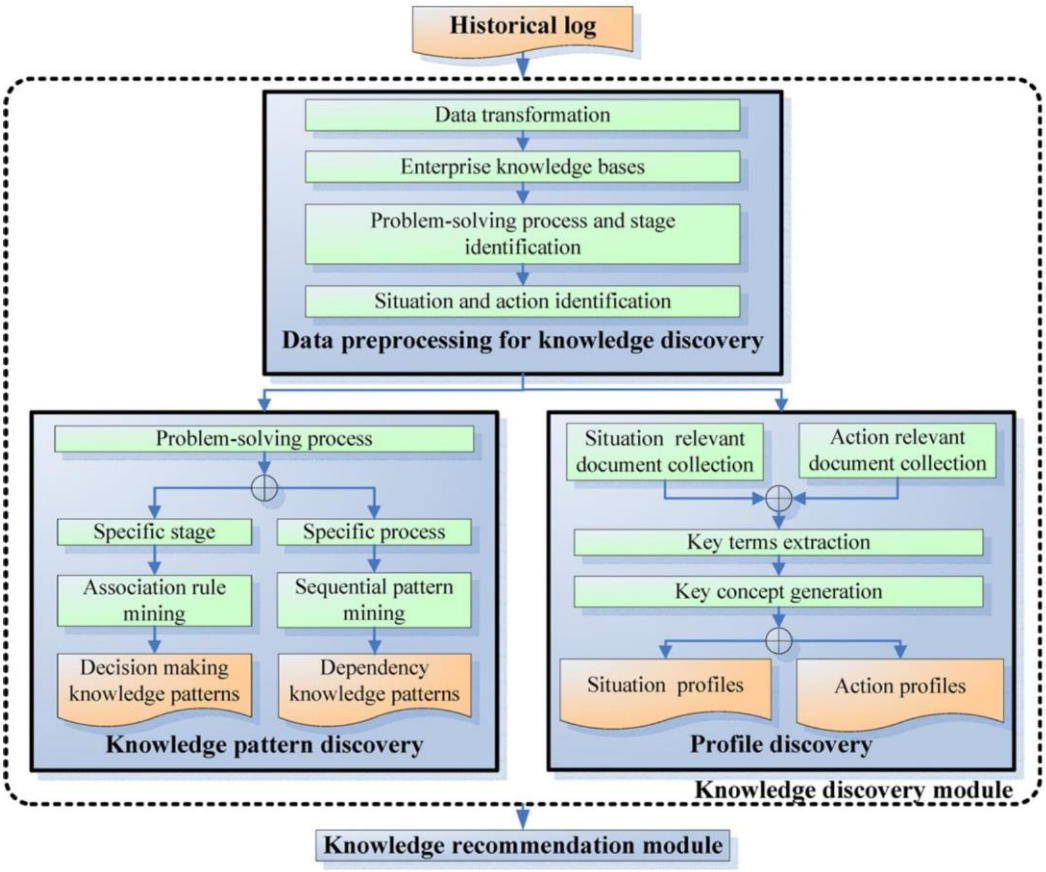

Figure 30: Procedure of knowledge discovery process[186]

Ur-Rahman and Harding present a general concept for mining text-based reviews and documents trying to establish text-mining also in the environment of production. Manufacturing systems provide textual data in various industrial contexts including product, service and quality but also data-related to project management. The proposed model detects key topics in the underlying textual data and classifies the documents in two categories aiming at decision support in the individual domains of production. In more detail the approach consists of three main layers as illustrated in Figure 31: The knowledge processing and storing unit including pre-processing, word representation by tf-idf and based on that the building of single key term phrases by

[185] cf. Liu/Ke, 2007, pp.147f
[186] Ibid., p.153

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

applying the k-means cluster algorithm. In the second layer Apriori Association Rule Mining is used for refining each cluster to Multiple Key Term Phrasal Knowledge sequences (MKTPKS) 3-termsets. This requires a relational database assigning each term of a cluster the corresponding document. The Apriori Association Rule Mining algorithm is then applied to each cluster, whereas the documents represent transactions, and terms represent items. Consequently, each document is characterized by the occurrence of three-term-sets gathered in a document term-set matrix. The third layer is dedicated to the classification using different methods such as Support Vector Machine or Naive Bayes. In advance to classification the data is manually labelled and inspected by domain experts dividing the document in a positive and a negative category. With the generated training set the algorithms learn to distinguish between good and bad information-containing documents. The obtained classification results finally showed that the concept of three-term-sets using the mentioned classification methods can significantly improve classification accuracy compared to single-term based approaches.[187]

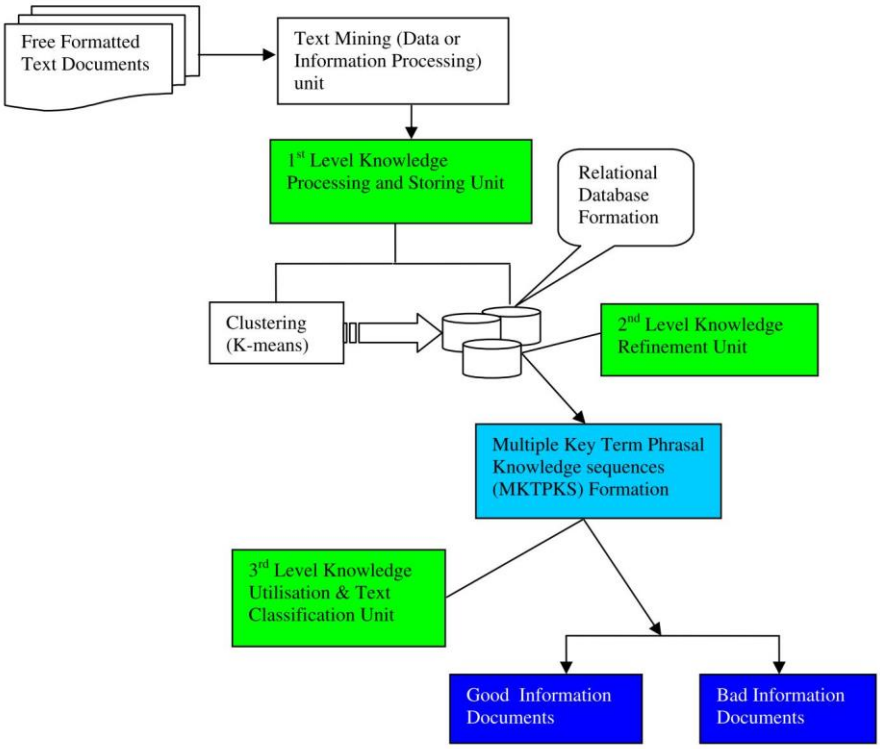

Figure 31: Architecture of MKTPKS-based knowledge management and classification system[188]

3.3 Text-Mining Approaches in Industrial Maintenance

In the area of industrial maintenance valuable textual information sources are maintenance logs which are manually drafted by maintenance staff for reporting.

[187] cf. Ur-Rahman/Harding, 2012, pp.4729f
[188] Ibid., p.4733

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

These logs implicitly contain information about equipment components and associated failures. The approach of Devaney et al. utilizes interdisciplinary methods of text-mining, unsupervised clustering and semantic representation for extracting that kind of information. They use the standards of Web Ontology Language (OWL) and Resource Description Framework (RDF) for building a semantic knowledge base. Thereby they don't learn ontologies but rely on existing ontologies being generally applicable across different types of machinery. Ontologies represent functions and interconnections of objects, systems and subsystems such as hydraulics, electronics, and pneumatics but also hoses, pumps and fluids. This underlying domain model builds natural categories which are utilized by cluster algorithms for categorizing maintenance logs to their functional groupings and for deriving component/repair or respectively component/action pairs. Additionally, text analytics is applied to correct spelling mistakes, replace abbreviations and detect higher level terms, which helps to reveal similarities between the records. With a case-based reasoning engine, solutions to occurring problems are found by analyzing similar past problems. It is based on four steps: Retrieval, ranking, recommendation, and confidence estimation, which finally provides a diagnostic troubleshooting suggestion.[189]

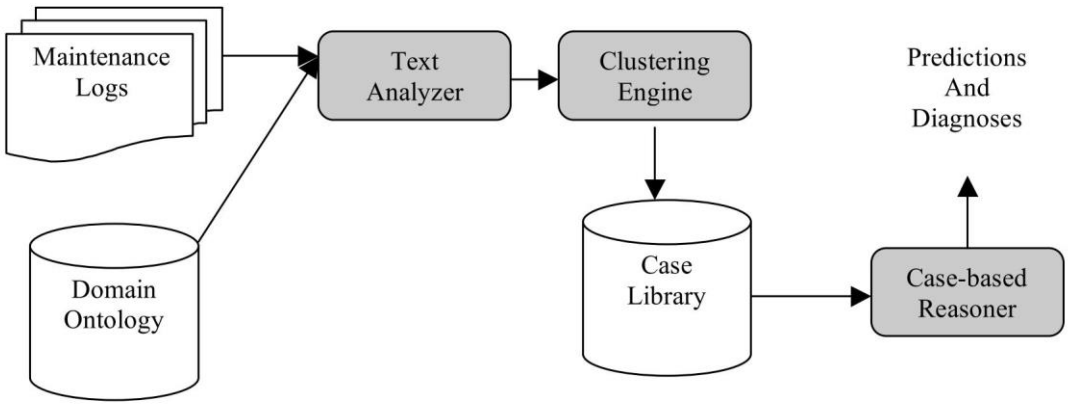

Figure 32: Model architecture for knowledge extraction from maintenance logs[190]

However, textual log data does not necessarily have to be manually drafted. Sipos et al. focus on logs recorded by applications running on equipment providing time stamp, event code and message of various occurring events. Their goal is to exploit that numerical and textual information for predicting failures based on a binary classifier. Since it's a supervised learning approach the data have to be labeled for identifying logs co-occurring with component failures. The authors avoid manual labeling by using service data listing necessary component replacements. Under the assumption of failure caused replacements the service data can be used for labeling the equipment logs. As a consequence, a target component can be assigned a set of

[189] cf. Ram/Devaney/Qiu/Lee, 2005, pp.1f
[190] Ibid.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

labeled historical equipment logs serving as foundation for failure prediction. Event codes and messages are taken as input features after converting them to numerical representation by the BoW concept. The daily logs are then gathered to bags representing for example a whole week. With the machine learning method of SVM correlations between the extracted features and the state of bags are learned which finally enables the state prediction of the current bag based on the features of current daily logs.[191]

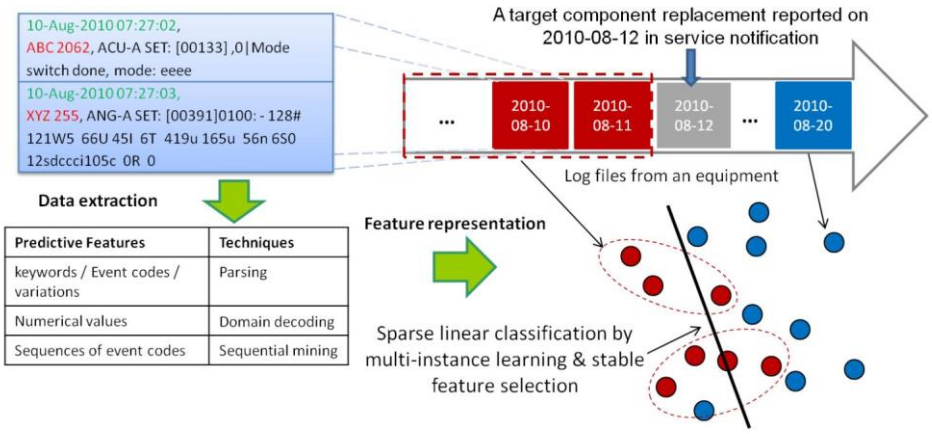

Figure 33: Overview of model architecture and methodology[192]

In the approach of Arif-Uz-Zaman et al. text-mining methods are applied to work orders and downtime data for extracting and analyzing failure time data. Work orders are an essential source since they provide information of the maintenance activities, whereas downtime data contains details of events whenever equipment was switched off. Merging these data bases now answers the decisive question whether a downtime event is caused by scheduled operations or unplanned events. With other words, down time data is used to label text fields of work order, classifying them in the categories of unplanned or non-failure. In the next step the raw text is prepared by various text cleaning methods including tokenization, stop words removal and stemming and is finally represented in a document-term matrix structure by BoW. For classification SVM is applied being an appropriate method especially for sparse instance vectors as they usually occur in text-mining tasks. The trained model finally enables the identification of urgent down time events based on corresponding work order descriptions.[193]

Finding the optimal maintenance strategy in terms of reliability, labor intensity, budgeting and spare parts logistics is the goal of McMillan's and Connor's text-mining and machine learning approach in industrial maintenance. They analyze work orders of a wind turbine company and build clusters according to their maintenance task.

[191] cf. Sipos/Fradkin/Moerchen/Wang, 2014, pp.1868f
[192] Ibid., p.1871
[193] cf. Arif-Uz-Zaman/Cholette/Ma/Karim, 2017, pp.388f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The outcome is a frequency distribution revealing the 40 most frequent groups of equivalent tasks providing a time-saving alternative to the manual analysis for supporting maintenance scheduling. The applied methods include standard procedures of data cleaning and spelling correction such as stop words removal and similarity measurement by Levenshtein metric as well as word vectorization converting words to a semantic vector space. With unsupervised machine learning methods such as k-means, similar vectors are found, and the obtained clusters finally evaluated and adjusted by experts. Structuring the outcome in further categories of components, plants or failure modes reveals additional information being a useful knowledge base for maintenance strategy selection.[194]

In recent industrial maintenance approaches text-mining methods also appear as submodules of comprehensive prescriptive maintenance approaches. Matyas et al. propose a procedural model for prescriptive maintenance in manufacturing industry involving multivariate data analysis as well as simulation tools for analyzing the historical data of product quality, machine failure, and production program. Together with incoming real-time machine data, the extracted correlations and patterns enable the prediction of system failures as well as the recommendation of prescriptive maintenance measures. Since the failure protocols, which represent a collection of historic maintenance measures, are available in semi-structured format containing free text passages, the authors also applied text-mining methods for knowledge exploitation. At first the protocols are cleansed and pre-processed for removing spelling mistakes and detecting synonyms. In the next step metadata is extracted for building new categories in order to exactly describe the affected part. As shown in the example of Figure 34**,** the algorithm identified "Module", "Assembly" and "Faulty part" as additional categories leading to a better structured and more precise failure description.

(A) Original format

Time-stamp	Machine ID	Problem	Countermeasure	Employee ID
13.01.17 12:03:01	AC01	Untypical noise at Spindle 1	Observation of Spindle 1, leaking sealing air, Air hose of connection ventile C was exchanged	Mechanic 01d

(B) Target format

Time-stamp	Machine ID	Module	Assembly	Faulty part	Employee ID	Repair time
13.01.17 12:03:01	AC01	Spindle 1	Spindle seal air	Air hose C	Mechanic 01d	1.2 h

Figure 34: Example of a failure protocol in original and target format[195]

[194] cf. Salo/McMillan, 2018, p.1
[195] Matyas/Nemeth/Kovacs/Glawar, 2017, p.462

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

The introduced concept of meta-data extraction aims at capturing, classifying and structuring data and is thus part of the first module "Data Acquisition and Pre-Processing". As illustrated in Figure 35, it is followed by the "Data Analysis & Simulation" which derives correlations of the previous structured data such as quality control measures and failure events. In the third element rules are defined and parameterized for real-time based prediction of quality deviations and machine failures. The element of "Prescriptive Maintenance Decision Support System" finally predicts system failures and recommends prescriptive maintenance measures based on the previous defined rules.[196]

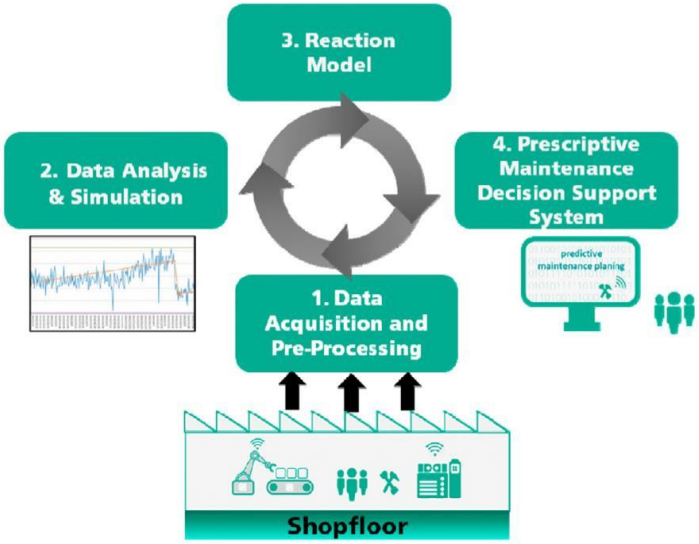

Figure 35: Procedural approach for prescriptive maintenance planning[197]

The PriMa model of Ansari et al. is another comprehensive prescriptive maintenance approach including text-mining methods for knowledge extraction from maintenance reports. Its key elements are the development and enrichment of capabilities for analyzing large amounts of multimodal and heterogenous data collected from multidimensional data sources as well as the generation of decision support measures and recommendations.[198] As shown in Figure 36, the proposed model consists of four layers, data management, predictive data analytic toolbox, recommender and decision-support dashboard and the semantic-based learning and reasoning overarching layer. The data management layer collects process-, product-, and machine-data building a multidimensional, multimodal and heterogenous data base. It considers horizontal and vertical data flow, which includes data flow within maintenance operation or maintenance management as well as semantic interlinking of operation and management data. Moreover, they are interlinked to a scalable data warehouse for handling multimodality. In the second layer multiple machine learning

[196] cf. Matyas/Nemeth/Kovacs/Glawar, 2017, pp.461f
[197] Ibid., p.462
[198] cf. Ansari/Glawar/Nemeth, 2019, p.1

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

and knowledge discovery algorithms are applied by systematic selection and combination based on the purpose of data analysis. The third layer correlates outcomes of analytic algorithms in order to prevent processing errors, to improve maintenance measures and finally to create recommendations. In the final overarching layer ontology and case-based reasoning are utilized to achieve a domain specific knowledge representation as well as to profit from solutions of similar past problems.

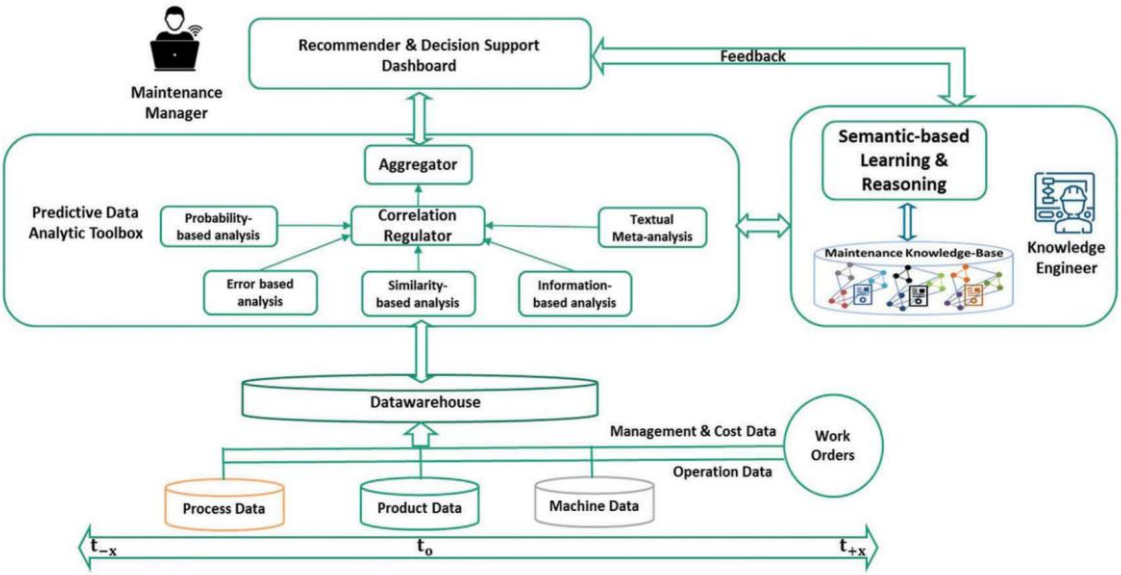

Figure 36: Overall model architecture[199]

In order to deal with maintenance reports of free text the authors propose a set of pre-processing methods as well as text-mining methods for extracting meaningful information. Figure 37 shows the procedure of text pre-processing followed by the CIMAWA concept of association measuring and AGF calculation. After splitting the documents to sentences, tokenizing refines them to individual tokens which are lowercased and finally compared to a stop words list for removing tokens of low information content. With the CIMAWA concept the asymmetric strength of association can be expressed based on the statistical measure of cooccurrence. The strength of attraction can then be measured with the AGF, revealing additional information of keyword relations in the provided maintenance text.[200]

[199] Ansari/Glawar/Nemeth, 2019, p.9
[200] cf. Ibid., pp.15f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

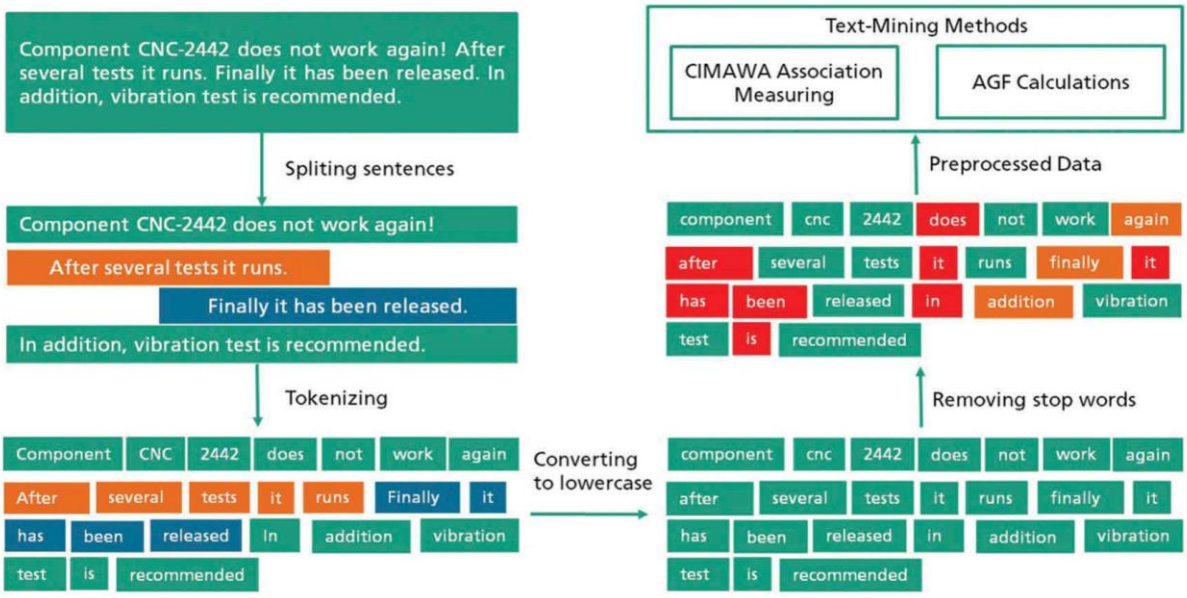

Figure 37: Example for pre-processing of free text[201]

Additional recent text-mining applications in industrial maintenance from the past five years which are also listed in Table 9 include the approach of Gunay et al. They analyze operator's work order descriptions for creating new benchmark system representing for example the component failure modes and breakdown occurrence frequencies by applying clustering techniques and association-rule mining.[202] Sexton et al. investigate the same type of source of textual maintenance work orders, however, focus on methodologies for knowledge representation and key word extraction building the foundation for further analysis. They utilize ontologies as concept for enabling machine readable knowledge representation as well as rule-based and data-driven tags methods for keyword extraction.[203] A recent approach of F. Ansari points out the lack of understanding maintenance text and proposes a model to quantify text on the semantic as well as syntactic level. The created text understanding map assists the maintenance planner based on three measures of text comprehension referred to as Association Measuring Index, Opinion Index and Cost Vector.[204] In the study of Wakiru et al. text-mining is applied for extracting the frequency distribution of conducted maintenance actions and affected equipment components along with their contribution to downtime loss from industrial maintenance reports. The extracted patterns serve as input of a discrete event simulation model for evaluating the effects of maintenance strategies on plant availability.[205]

[201] Ansari/Glawar/Nemeth, 2019, p.15
[202] cf. Gunay/Shen/Yang, 2019, pp.518f
[203] cf. Sexton/Hodkiewicz/Brundage/Smoker, 2018, pp.1f
[204] cf. Ansari, 2020, pp.3f
[205] cf. Wakiru/Pintelon/Muchiri/Chemweno, 2018, pp.899f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

3.4 Summary of the literature review

The following Table 9 shows an overview of the presented approaches with respect to application area, objective and implemented methods. First of all, text-mining has shown to be successfully applied across various sectors involving fields such as marketing, customer intelligence, finance, healthcare as well as the industrial sector. So far, text analysis predominantly aims at additional input generation for machine learning algorithms in order to predict future events or to find relevant groups of interest by clustering. Direct exploitation and knowledge visualization are mainly limited to sectors of marketing and customer intelligence by drawing benefits from opinion mining techniques and sentiment analysis. Focusing on the domain of industrial maintenance, apart from the mentioned novel approaches of prescriptive maintenance, failure prediction as well as document retrieval for problem solving are the prevalent approaches for text utilization. Although the mentioned concepts have improved decision-making and troubleshooting, this work tries to enhance existing models by involving a new approach for knowledge extraction and visualization. Applying sentiment analysis also to the area of industrial maintenance and combining it with the concept of association measuring enables to profit from the human background-knowledge implicitly hidden in text. It creates a visualized knowledge base for decision makers by deriving inferences and opinions with respect to entities, actions and events being not limited to the task of failure prediction. Thus, this work can be seen as attempt to close the existing gap in literature by providing a procedural model for textual knowledge exploitation for the sector of industrial maintenance.

Table 9: State-of-the-art content overview

Nr.	Field of application	Objective	Method
1	Customer relationship mgt.	Customer churn prediction	SVD/Neural network
2	Marketing/Advertising	User group detection	Sentiment Analysis
3	Healthcare	Injury root cause detection	Association Measuring
4	Construction project planning	Cost-overrun prediction	SVD/Radial Basis Functions
5	Finance/Investment	Stock market prediction	tf-idf/SVM
6	Healthcare	Identification of complications	String search/Keywords
7	Production/Routing	Identification of generic routings	Similarity measuring
8	Production/Troubleshooting	Specific knowledge retrieval	Association rule mining
9	Production management	Document classification/Support	SVM /Association rules

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

10	Industrial Maintenance	Diagnostic troubleshooting	OWL/CBR
11	Industrial Maintenance	Equipment state prediction	SVM/BoW
12	Industrial Maintenance	Classification of down time events	SVM/BoW
13	Industrial Maintenance	Maintenance scheduling support	k-means/Word2Vec
14	Industrial Maintenance	Prescriptive decision support	Data analysis/Simulation
15	Industrial Maintenance	Prescriptive decision support	Semantic CBR/Data mining
16	Industrial Maintenance	Benchmark development	Clustering/Rule mining
17	Industrial Maintenance	Knowledge representation	Ontologies/Keywords
18	Industrial Maintenance	Maintenance text comprehension	Semantic/Syntactic mining
19	Industrial Maintenance	Maintenance strategy simulation	Discrete sim. /Keywords

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

4 Applied Research Design Methodology

4.1 Model Development with Design Science Methodology

Model development for use-case application is based on guidelines for research in information systems by Alan R. Hevner[206]. Its fundamental principle focusses on the scientific creation of models for understanding and solving design problems as follows:

1. Design an artifact: Building of a construct, model, method or an instantiation
2. Problem relevance: Development of technology-based solutions to important and relevant business problems
3. Design evaluation: Evaluation of the design artifact by evaluation methods
4. Research contributions: Providing clear and verifiable contributions in the areas of the design artifact.
5. Research rigor: Application of rigorous methods in the construction and evaluation of the design artifact
6. Design as a search process: The search for an effective artifact requires utilizing available means while satisfying laws in the problem environment
7. Communication of research: Presentation to technology-oriented as well as management-oriented audiences.[207]

Design science is basically a process of finding appropriate solutions for problems. The fundamental principle of design-science research from which these seven guidelines are derived is that knowledge and understanding of a design problem and its solution are acquired in the building and application of an artifact.

The first guideline emphasizes the requirement of innovative and purposeful artifacts addressing an important organizational problem. In this approach the term artifact is defined in a broad sense and involves not only instances but also the constructs, models and methods applied in practical implementation of information systems. On the other side it does not include people and the organizational and social contexts in which artifacts are used. They are seen as interdependent and coequal with artifacts in business needs.[208]

[206] cf. Hevner/Salvatore/Ram/Park, 2004, p.82
[207] cf. Ibid., p.83
[208] cf. Ibid., p.82

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Problem relevance is the main issue of the second guideline. In the field of information systems and artifact development research aims at knowledge acquisition enabling the implementation of technology-based solutions of yet unsolved business problems. In contrast to behavioral science, which generates theories for understanding outstanding phenomena, design science focuses on changing the phenomena by innovative artifacts. For instance, technology acceptance is a common and relevant business issue being predicted and explained by theories of behavioral science. Based on that design science's goal is to develop artifacts in order to cope with these acceptance challenges.[209]

The quality and utility of an artifact has to be assessed by evaluation methods, which is focused on in the third guideline. Artifact evaluation is strongly linked to the business environment and its arising problems the artifact deals with building the requirements and thus the foundation for evaluation. It prerequisites the definition of appropriate metrics as well as the gathering of relevant data for measuring predefined attributes. An artifact can be evaluated in many different aspects including functionality, completeness, consistency, accuracy, performance, reliability, usability or fit with the organization. Since design is an iterative process the evaluation should not only be considered as an act of final assessment, it also provides valuable feedback for quality improvement during construction phase. A list of methods for conducting evaluation can be summarized as follows:[210]

Table 10: Evaluation methods and categories

1. Observational	Case Study: Study artifact in depth in business environment
	Field Study: Monitor use of artifact in multiple projects
2. Analytical	Static Analysis: Examine structure of artifact for static qualities (complexity)
	Architecture Analysis: Study fit of artifact into technical IS architecture
	Optimization: Provide optimality bounds on artifact behavior
	Dynamic Analysis: Study artifact in use for dynamic qualities (performance)
3. Experimental	Controlled Experiment: Study artifact in controlled environment for qualities
	Simulation - Execute artifact with artificial data
4. Testing	Functional (Black Box) Testing: Execute artifact interfaces to discover failures
	Structural (White Box) Testing: Perform coverage testing of some metric in the artifact implementation
5. Descriptive	Informed Argument: Use information from the knowledge base (relevant research) to build a convincing argument for the artifact's utility
	Scenarios: Construct detailed scenarios for utility demonstration of the artifact

Guideline four focuses on the perspective of research and the ultimate question: What are the new and interesting contributions? There are three main areas in which the contributions in the environment of design science can be classified. The first group involves new insights concerning the design artifact itself. As stated in the

[209] cf. Hevner/Salvatore/Ram/Park, 2004, p.84
[210] cf. Ibid., pp.85f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

second guideline, the artifact should represent a technology-based solution for a relevant and yet unsolved business problems. It may extend existing knowledge or apply existing knowledge in new innovative approaches both contributing to design science research. The second category is not dedicated to the problem itself, but to its foundations. This involves improvements and inventions of constructs, models, methods and instantiations. Modeling formalisms, ontologies, problem and solution representations and design algorithms are examples belonging to that category. Finally, the development of new evaluation methods and metrics for experimental, analytical, observational, testing or descriptive evaluation is a form of research contribution.[211]

The research rigor mentioned in the fifth guideline focuses on the way the research is conducted. Design-science requires rigorous methods in construction and evaluation which is often linked to mathematical formalism describing the specified artifact. Although mathematical rigor is an important part of research, assessment should also be conducted with respect to applicability and generalizability avoiding over-abstraction and omission of important parts of the problem. Thus, research rigor in design science is derived from the effective use of the knowledge base built by the theoretical foundations and research methodologies.[212]

Design science is an iterative process aiming at discovering solutions to relevant business problems being represented by an interaction between design generation and the testing of alternatives. As outlined in the sixth guideline, the search for an effective solution can be viewed as utilization of available means while satisfying restrictions of the existing environment. Means involve the set of actions as well as resources available for artifact construction. Thus, successful problem solving requires knowledge in both, the application domain defining the constraints, and the solution domain including technical and organizational issues.[213]

Finally, the artifact and the results of research must be presented to technology oriented as well as management oriented audiences. Communication of results to technology-oriented audiences focuses on detailed artifact description for clarifying construction and application. The main goal is to enable the understanding of the artifact and its development being essential for practitioners, in order to benefit from the provided solution, but also for researchers obtaining a knowledge base for further extension. Management oriented communication emphasizes the importance of the problem and the usefulness as well as effectiveness of the provided solution. Moreover, the required resources for implementation and application should be

[211] cf. Hevner/Salvatore/Ram/Park, 2004, p.87
[212] cf. Ibid., pp.87f
[213] cf. Ibid., p.88

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

communicated. Thus, the focus is less on the functionality of the artifact itself than on the problem relevance and the knowledge required for application.[214]

4.2 Statistical Evaluation

4.2.1 Classification Evaluation

A common tool for assessing classification accuracy is the **confusion or error matrix**. It is a square array of dimension $n \, x \, n$ where n is the number of classes. The rows of the matrix visualize the predicted classification of instances the columns represent the real conditions. The following table shows the confusion matrix for binary classification:

Table 11: Confusion matrix[215]

Total population	Condition positive	Condition negative
Predicted condition positive	True positive	False positive
Predicted condition negative	False negative	True negative

Based on these error categories the measurements recall and precision can be calculated. Recall is the proportion of correctly predicted positive cases over the total amount of real positive cases. It can be interpreted as measurement for completeness and is despite its meaningfulness tended to be neglected in machine learning applications.[216] Conversely, precision or confidence represents the proportion of correctly positive cases over the total amount of predicted positive cases. It evaluates the exactness and is thus an important quality measure. Being TN the number of true negative, TP true positive, FN false negative and FP false positive assignments precision and recall can be calculated as follows:[217]

$$Precision = \frac{TP}{TP + FP}$$

Equation 20: Formula of Precision

$$Recall = \frac{TP}{TP + FN}$$ [218]

Equation 21: Formula of Recall

Combining precision and recall leads to the **F1-score** striking balance between those measures. F1-score is calculated by the harmonic mean of precision and recall and ranges from zero to one. A score of zero means either precision or recall are zero, a score of one means 100 percent precision and recall.

[214] cf. Hevner/Salvatore/Ram/Park, 2004, p.90
[215] cf. Fawcett, 2006, p.862
[216] cf. Powers, 2011, p.38
[217] cf. Ibid., p.38
[218] cf. Fawcett, 2006, p.862

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$ [219]

Equation 22: F1-score formula

The **Matthews correlation coefficient** is a measure for evaluating binary classification and is commonly used for machine learning algorithms. It's a value in the interval [-1,1] whereas 1 indicates perfect and 0 random classification. It measures the correlation between the predicted classifications and the target values by considering all four different error types. Thus, a positive coefficient implies a positive correlation between the prediction and the target values. The Matthews' correlation is given as follows:

$$MCC = \frac{TN * TP - FN * FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$ [220]

Equation 23: Matthews correlation coefficient

Kohen's Kappa is a metric invented for measuring the degree of agreement and disagreement of two or more people observing the same phenomenon. This method is also applied to derive the accordance of model's predictions and reality or to measure the extent of agreement between a classification set and some expert system. In contrast to conventional accuracy measuring, Cohen's kappa takes into account the portion of the counts that may be attributed by chance and subtracts them. Thus, Cohen's Kappa can be expressed by the observed accuracy $p_O$and expected accuracy p_e and is defined as follows:[221]

$$\kappa = \frac{p_O - p_e}{1 - p_e}$$

Equation 24: Cohen's kappa coefficient

Using the introduced definition of the confusion matrix the observed and expected accuracy turn into:

$$p_O = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 25: Formula of observed accuracy

$$p_e = \frac{(TN + FP) * (TN + FN) + (FN + TP) * (FP + TP)}{(TP + TN + FP + FN)^2}$$ [222]

Equation 26: Formula of expected accuracy

[219] cf. Nicolas, 2014, p.69
[220] cf. Măndoiu/Zelikovsky, 2007, p.412
[221] cf. Ben-David, 2008, p.826
[222] cf. Ibid., p.827

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

4.2.2 Sample Selection and Sample Size Definition

Since the number of classified instances is usually very big, it would be impractical to investigate the whole population. Thus, to estimate the prediction accuracy a sample is drawn by selecting a finite number of units from the population. The goal is to estimate characteristics of the population such as a mean or a proportion after only observing the sample.

There are two general types of sampling, probability sampling and non-probability sampling. In probability sampling all instances of the population have an equal and known probability for being selected. One representative of that type of methods is simple random sampling using a number table for enumerating the population and selecting randomly a desired number of units. It provides good results in terms of generalizability but can be more work intensive than non-probability approaches. In contrast to that, population instances in non-probability sampling don't have equal chances for being selected and their probabilities are in general unknown. Sampling obliges the analyst and his or her subjective judgement biasing the inference from the sample to the population. Non-probability is thus used for generating hypotheses instead of hypothesis testing.[223]

An important factor influencing the representativeness and the ability to make inferences from samples to the population is the sample size. On the one hand, it has to be large enough to be of practical significance; on the other hand, too large samples come along with unnecessary costs and effort. A common method for deriving an appropriate sample size is the assessment based on a confidence interval. Since accuracy measuring is based on error categories, in this application confidence intervals for proportions are regarded. The estimator of a proportion is

$$\hat{p} = \frac{X}{n}$$

Equation 27: Estimator of a proportion

whereas X is are the considered instances of a category and n the sample size. The estimated proportion has a binomial distribution but can be approximated by the normal distribution for large enough n. The confidence interval with a defined significance level α and the corresponding z-value of standard normal distribution can then be calculated by:

$$(\hat{p} - z * \frac{\sigma}{\sqrt{n}}, \hat{p} - z * \frac{\sigma}{\sqrt{n}})$$

Equation 28: Confidence interval based on significance level

[223] cf. Morais/Barragués/Guisasola Aranzabal, 2014, p.178

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Using the Wald method, the variance σ^2 is substituted by $\hat{p} * (1 - \hat{p})$. the width of the interval W equals:

$$W = 2 * z * \sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}}$$

Equation 29: Width of confidence interval

Defining $h = W/2$ as the half-width the sample size n is given by:

$$n = \frac{z^2 * \hat{p} * (1 - \hat{p})}{h^2}$$

Equation 30: Estimated sample size based on the confidence interval

The required half-width and the significance defines z and h, the proportion has to be estimated for example by using a small sample size at the beginning. Sometimes the variance is set to its maximum at $\hat{p} = 0.5$ and $\hat{p} * (1 - \hat{p}) = 0.25$, however, it is more cost effective to start with a smaller sample size and roughly estimate the order of $\hat{p}$, since the proportion has a large direct effect on the sample size.[224]

[224] cf. Newcombe, 2013, p.24

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

5 Design and Implementation of Procedural Model for Knowledge Extraction from Textual Maintenance Records

5.1 Use-Case Description

Company A is an industrial manufacturer using technically advanced equipment and processes for production. Notably, the name of the company and sector is not mentioned due to data privacy reasons. In that technology-oriented business maintenance activities play an important role for ensuring equipment availability and manufacturing continuity which consequently has a significant impact on company's total profit. The maintenance activities of service, inspection, repair and improvement are conducted and reported by company-internal maintenance staff being familiar with the specific domain. The reporting is based on a table structured in affected equipment, text of the graphical user interface (GUI), down event, conducted maintenance action, the author of the report and the time stamp. As shown in Table 12, apart from the provided meta-data the reports are unstructured and are drafted by the maintenance workers without any limitations or reporting suggestions. As a result, the style, phrasing, correctness and level of detail of the maintenance logs strongly depend on the author. Especially in the column "Aktion", describing the conducted maintenance action, reports differ due to the freedom in drafting. The column "GUI_Buchungstext" provides information about the occurred error and is either drafted by maintenance staff or automatically generated from error logs. For ensuring data privacy the names and initials of the authors are replaced by "Person" and "x".

Table 12: Example of maintenance records

EQUIPMENT	GUI_BUCHUNGSTEXT	DOWN_EVENT	AKTION	AUTHOR
L2301	APCFDC: T_PROC_ME_V_F , VE814692, L2301_PM2, ETCH TIME MAIN ETCH; ETCH TIME DEVIATION	ST	APC Daten überprüft... ok xx	Person1
L2303	xxxxx keine horden erkennung	ST	Anlagenrechner neu gestartet => o.E.	Person2
L2303	xxxx:Transport EdcID:20160205021717	ST	Wafer entladen... xxxxx	Person3
L2302	APCFDC: T_ME_V EQUIPMENT:L2302_PM1,LOT:VF745533, RECIPE:x_xx_xx, ANLAGE AN IH; LOS AN SL AETZUNG; ETCH TIME DEVIATION	ST	APC Alarm wegen Prozessabbruch aufgetreten, Wafer ausgeladen > ok xx	Person4
L2301	xxxxxx Counter	ST	Burn gefahren -> emeute Drucküberprüfung erforderlich xx	Person5

Apart from documenting equipment failures and maintenance actions, the reports contain assessments, requirements and recommendations which are based on the expertise of the maintenance staff and usually mentioned in the column "Aktion". Due

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

to that domain knowledge and latent coherences hidden in text, maintenance logs are regarded as a valuable source for decision support and improvement. However, there are about 10,000 maintenance records stored, which in full extent are not manageable and inappropriate for manual analysis. Since there is currently no corresponding software in use, the maintenance management cannot profit from that textual data source.

The proposed procedural model described in the following section aims at providing a sophisticated framework for that problem by enabling knowledge, extraction from text. Because of the domain specific terms and knowledge, the visualized output is provided to experts and should assist them in drawing conclusions and making decisions.
Moreover, the approach does not only focus on the extraction, it should also reveal weak spots in reporting and detect knowledge gaps which might be closed by an improved standard reporting structure.
The following chapter is structured as follows: At first, the general architecture and its main components are explained trying to give an overview and describing the occurring challenges. In the next step, the implementation is elaborated in more detail by referring to the tasks of pre-processing, word embedding keyword extraction, association measuring and sentiment analysis and applying the already discussed theory of Chapter 2 to the provided data. Finally, the results are visualized and discussed with respect to the company related problem and the research question.

5.2 Overview of Procedural Model for Knowledge Extraction

In Figure 38 the whole architecture of the procedural model for extracting knowledge from text is illustrated. To begin with the lower left corner, maintenance staff is informed about the required maintenance action and accomplishes them according to the maintenance schedule. As already shown in Table 12, the next action of reporting is based on predefined meta-data, providing a rough structure for documentation. Maintenance workers record events and attributes such as the conducted action, name, down event and error type as well as further actions required if the procedure has been unsuccessful. The drafted maintenance reports are stored in an excel file forming the textual data base with currently about 10,000 reports of ten different plants. The data is categorized in six columns, of which “Aktion” and “GUI_Buchungstext” are composed of free text and considered to carry meaningful but also heavily exploitable information.

Before the analysis can be started the data has to be pre-processed enabling the computational accessibility. This includes quality checks for example the cleansing of

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

unprocessable characters such as arrows, the correction of spelling mistakes, the detection and substitution of abbreviations and the recognition of synonyms. In this use-case pre-processing represents an essential part of the procedural model due to the lack of reporting structure and individual preferences of authors in drafting. Moreover, the ambition to save time and to conduct reporting as efficiently as possible, leads to the vast use of abbreviations and the disregarding of grammatical rules, which makes pre-processing indispensable. In that model pre-processing also involves the determination of part-of-speech tags for labelling the words according to their grammatical properties. That task is necessary for finding the root form of words by lemmatization but also for extracting the key words in single reports. The next step of word embedding converts text to numbers using the term-frequency inverse-document-frequency (tf-idf) representation. It is a measure quantifying the importance of words based on the frequency of pre-processed words in text. The keyword extraction is based on the previously determined part-of-speech tags by combining a noun and a verb to an individual phrase. Since the reports are in the most cases kept short consisting on average of 38 characters omitting conjunctions, complex sentence structures and even objects, the extraction of the noun and the verb often corresponds to the subject and predicate of the sentence building a satisfying representation of the maintenance report.

Combining the approaches of association measuring and sentiment analysis is the main idea behind the model's analysis section providing an extension to the already mentioned approaches (cf. Section 2.2). At first the associations between extracted key phrases are calculated based on an adaption of the introduced CIMAWA algorithm. To mention one example, the association between the action "Wafer entladen" and the error "Temperatur" can be calculated imitating the human capability of asymmetric inference which builds finally the foundation for quantifying their gravitational attraction. Apart from the associations, the sentiment of the reports and the corresponding key phrases are determined drawing on opinion word dictionaries. With these lexical sources the pre-processed words and consequently entire reports are classified to sentiment categories. Again, the procedure has to be adapted to the environment of maintenance requiring the specification of dictionaries and creating rules for classification. This includes words such as "required" or in German "erforderlich" indicating unsuccessful troubleshooting and the demand of additional actions, which represents a negative sentiment in that special domain. Exclusively relying on conventional dictionaries would lead to biased results classifying the mentioned example as neutral. Since reports mainly report errors and actions, positive classification is obsolete, resulting in a binary concept with neutral and negative categories. Based on that approach each keyword phrase is assigned a label according to the percentage of the co-occurring sentiment depicted by the color shades from green to red.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

The associated and colored word phrases build a network represented to the maintenance manager, who is familiar with the domain and who is able to derive conclusions as well as further actions. This includes tasks such as maintenance scheduling but also template adaptions for improving data quality and closing knowledge gaps.

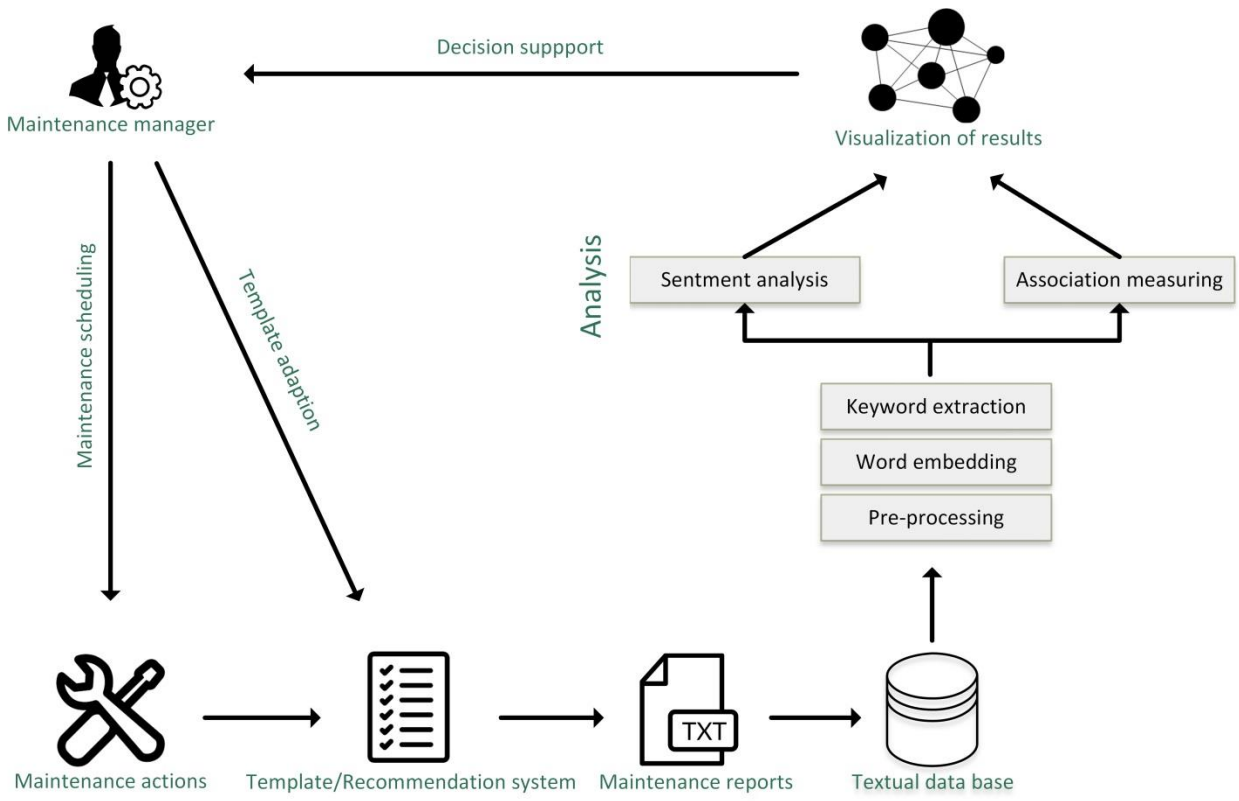

Figure 38: Procedural model for knowledge extraction

5.2.1 Pre-processing, Word embedding and Keyword extraction

Figure 39 lists the applied methods and tasks in their sequential arrangement. Obviously, the required number of procedural steps for pre-processing stands out compared to the remaining tasks, which again emphasizes its importance and extend in the underlying use-case. In the column of keyword extraction, the last connection is represented by a dotted arrow indicating its optional character. The conditions for the optional branching are explained in the corresponding section of keyword extraction.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.

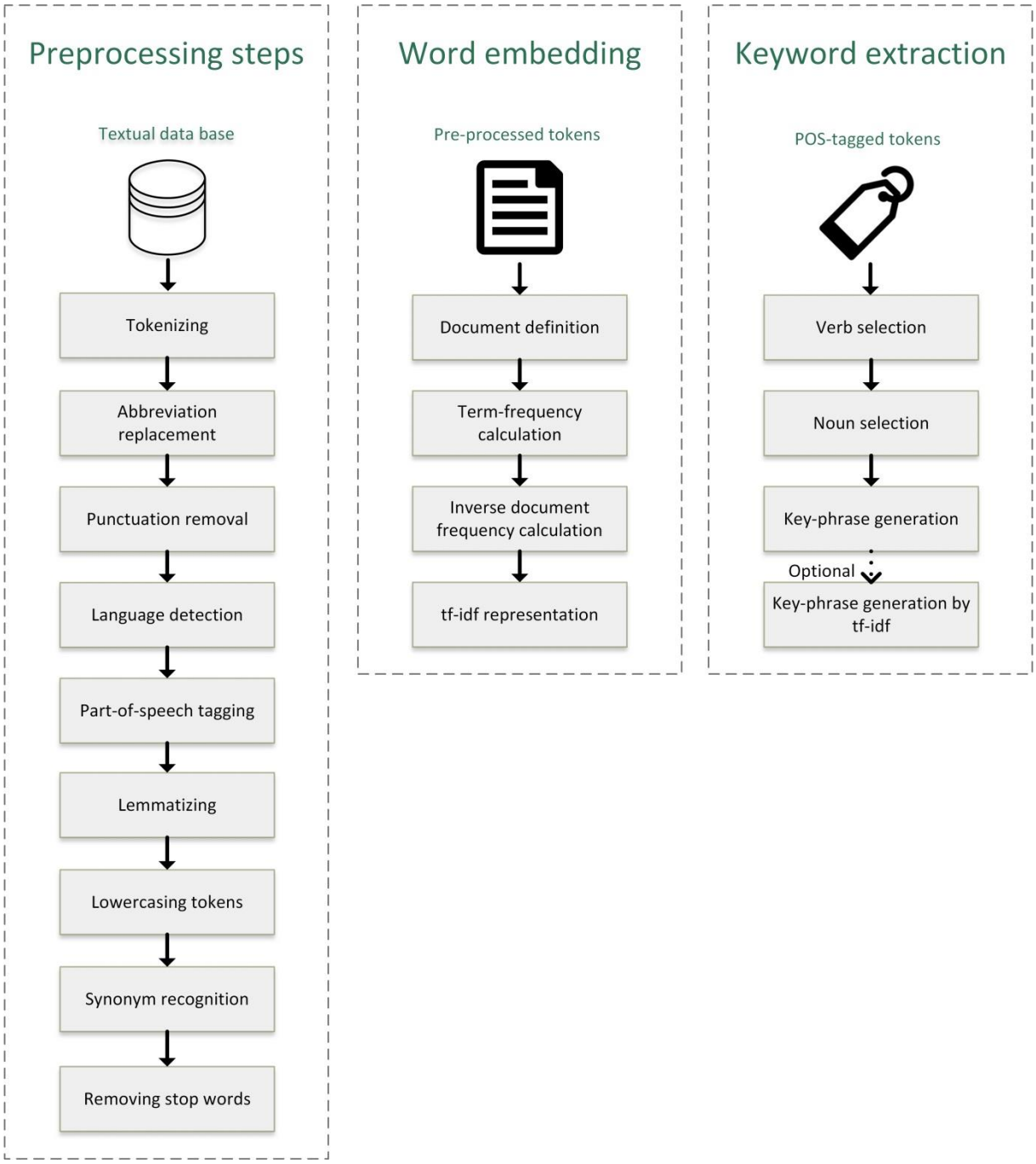

Figure 39: Steps of pre-processing, word embedding and keyword extraction

a.) Pre-processing

The pre-processing model is explained by successively applying the methods to a report example and showing the outcome of each step. For accomplishing tokenizing, English POS-tagging and stop words removal the created software prototype draws on the Natural Language Toolkit for Python providing corpora and lexical resources.[225] Pre-processing of German tokens requires modifications and different lexical sources, which are listed in the corresponding section.

[225] cf. Bird/Klein/Loper, 2009, pp.181f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

At first the whole report is split into single tokens based on space delimiters.

Figure 40: Tokenizing applied to maintenance reports

The abbreviation o.E means “ohne Erfolg” translated to English “unsuccessful” which has to be replaced by the full terms. This is essential since it indicates a negative sentiment which could not be extracted from the abbreviation. Due to the reuse of typical abbreviations, occurring in a manageable amount, a dictionary was generated assigning them the full terms. In case the replacement of the abbreviation hasn’t any impact on the analysis and doesn’t enrich the information content, the replacement is omitted. One example for that is “ACP” meaning “Automatic Process Control” which does not add any valuable information regarding the following analysis. Conducting the step of abbreviation removal leads to:

Figure 41: Example of abbreviation replacement

Strikingly the arrow sign has to be removed along with other non-processable characters such as question marks or hyphens.

Figure 42: Example of punctuation removal

Since reports are sometimes drafted in English and the following part-of-speech tagging depends on the underlying language, an annotator based on a Naïve Bayesian Classification is used for distinction. For implementation the tool “langdetect” from Google’s language detection library is used providing the annotation of 55 different languages. It is based on a probabilistic approach deriving language probabilities from the underlying word sequence.[226] The example sentence shows the highest probability for the German language.

[226] cf. URL: https://code.google.com/archive/p/language-detection/, 03.09.2019

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

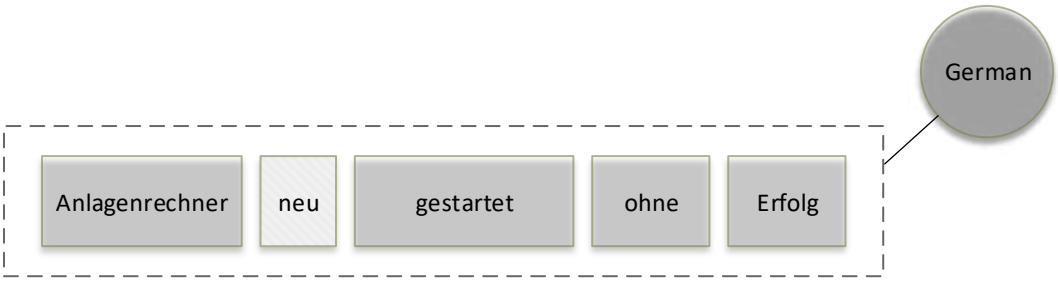

Figure 43: Language identification of maintenance reports

Based on the language annotation either the English or the German based recurrent neural networks are selected for accomplishing part-of-speech-tagging. Each network was trained on language specific labelled corpora classifying the tokens in lexical categories. For German POS-tagging the TIGER corpus of the University of Stuttgart is used as training data set consisting of 900,000 POS-tagged tokens.[227] The neural network achieved an accuracy of 96.09% after being trained on 90% of the Corpus and extends the NLTK classifier based tagger with a modified feature detector. English POS-tagging is achieved by the standard NLTK toolkit which was trained on NLTK corpora.[228]

The provided example is divided in four different word classes, *NN: noun, ADJD: adjective, VVPP: past participle of verb, APPR: left hand preposition*. Obviously, the conducted method provides very detailed tags distinguishing between different types of prepositions based on the word sequence or different noun classes, such as entity names and adjectives used as nouns. Since the reports are usually not framed in full sentences, the tagging network does not always produce correct results with respect to that detailed classification. However, for conducting the following analysis the assignment to the main categories, noun, verb, adjective and preposition is sufficient, and the method can be applied with a sufficient accuracy.

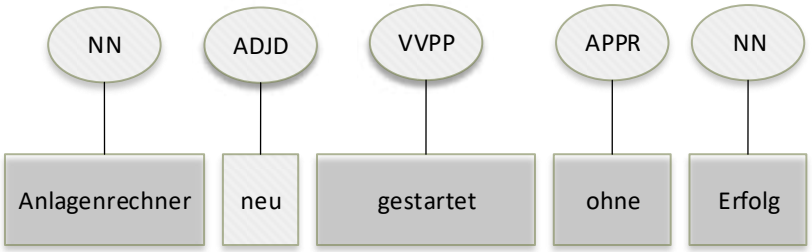

Figure 44: POS-tagging of maintenance reports

Providing the additional morphological information of part-of-speech enables the pre-processing step of lemmatizing, which extracts the root form of a token. This is essential for algorithms relying on word frequencies such as tf-idf or association measuring since it minimizes the number of individual tokens by equalizing word

[227] cf. URL: https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html, 09.09.2019
[228] cf. Bird/Klein/Loper, 2009, pp.203f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

inflections, pluralization and declinations. The applied method is again based on the TIGER corpus or respectively on the NLTK toolbox. The sequential recurrent neural networks predicting the lemma of each word transform the past participle of the example sentence to the infinitive form leading to:

Figure 45: Lemmatizing of maintenance reports

In the final steps the tokens are lowercased, and verbs are checked for synonyms based on a dictionary. Referring to Table 12 row three and four, the derived infinitives “ausladen” and “entladen” have the same meaning and are recognized as analogously equal verbs. It is important that the lemmatization is applied first, as the used synonym dictionary is based on infinitives. In the last step, the tokens which do not add any significant information to the sentence are removed; also referred to as stop words. There lists of stop words available for different languages of texts generated by statistical analysis of information content. Nevertheless, the list has to be adapted to the individual application. For example, in this use-case the stop word “nicht” is removed from the stop-word list since it has an impact on the sentiment and on the final output.

The proposed steps are designed for reports in the column “Aktion” or “GUI Buchungstext” being manually drafted by maintenance workers. Automatically generated logs show a different structure, and do not require morphological based pre-processing. Since logs are composed in a coherent style providing information in a predefined structure, rule-based approaches are proposed to be suitable for knowledge extraction. Regarding the first example in Table 12, the key information about the occurred error is found in the last part of the log, namely “etch time deviation”. This means, after detecting the logs by their capital letters and underscores the key words are extracted based on the position which can be implemented with a simple if- clause.

b.) Word representation and Keyword extraction

Word embedding covers all methods striving for the conversion of text into a numerical representation of real numbers. In this procedural model, the term frequency – inverse document frequency (tf-idf) is used to fulfil that task expressing the importance and meaningfulness of a word. Compared to other word embedding methods it stands out by its simplicity and its rather easy implementation. Moreover, in contrast to the in Section 2.2.2 introduced Word2Vec model tf-idf does not depend on context words and is thus less affected by incomplete and insufficient reporting.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

In that approach the proposed numerical conversion has two purposes. Firstly, it serves as a second instance for keyword extraction, whenever the preferred method of verb+noun phrase generation is unsuccessful. This can be the case for example if reports are just single words and there can't be found a noun+verb phrase in the report. The two words with the highest tf-idf values are then selected as key words. Secondly, it is an essential part in the concepts and formulas of association measuring and AGF, by expressing the importance of already extracted key phrases.

In the first stage of tf-idf the documents have to be defined which should be not confused with the general linguistics meaning of the term. Documents are text segments and can also be paragraphs or single sentences. In this use case each report represents a new document and the tf value for each word in that single report can be calculated as follows:

$$tf(t,d) = \frac{f_{t,d}}{n_d}$$ [229]

Equation 31: Implemented term frequency concept

The term frequency equals the frequency of the term t in document d $f_{t,d}$ divided by the number of words in document d n_d. This means the more often words appear in a report the higher the tf value, whereas long reports diminish the score. Applying this formula to the extracted key phrases simplifies it to the reciprocal of n_d since key phrases occur only once in each report. Taking into account the appearance of words or phrases across all provided documents leads to the idf formula:

$$idf(t,D) = \log\left(\frac{N}{|d \in D: t \in d|}\right)$$ [230]

Equation 32: Implemented inverse document frequency concept

The base-10 logarithm is taken from the total number of documents N divided by the number of documents in which the regarded term t or respectively the regarded key phrase occurs. This means words which are used very frequently being present in many reports come along with a small idf value since the fraction within the logarithm is then close to 1. Finally, the multiplication $tf_idf = tf * idf$ results in the tf-idf score representing the word or key phrase importance.

In the following model illustration, the procedure of key word extraction and tf-idf representation is explained by applying the steps to the pre-processed example from above. The noun and verb of the report are selected as key words building a representative key phrase. If there are more than one noun or verb tags the first noun and the first verb is extracted. Although it is possible to permute the sentence clauses

[229] cf. Manning/Raghavan/Schütze, 2008, p.117
[230] cf. Ibid., p.118

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

in the German language, the first noun is likely to be the subject or part of the subject since sentences are short and subjects are preferably put in the first position. The existence of more than one verb could appear in combination with conjunctions or passive constructions, however, in this use-case auxiliary verbs are almost always omitted and the existence of more than one verb per report is very scarce. Therefore, the circumstance of multiple verbs is not explicitly considered in this approach. The goal of key word extraction is to describe the conducted action and the corresponding subject of each report for building the foundation of association measuring while keeping the model as simple as possible. At this stage the adjectives and equipment states are not relevant yet and will be of concern in the section of sentiment analysis. Therefore, the extraction of a phrase consisting of a noun and a verb serves as a satisfying construct, whereas the proposed method can be seen as an approximation for subject and predicate extraction building the fundamental parts of a sentence.

After extracting the key phrase, the tf-idf value is calculated. In this example $f_{t,d} = 1$ and $n_d = 5$ which results in $tf(t,d) = 0.2$. The key phrase occurs 82 times in the data base which results in $idf(t,D) = \log\left(\frac{9225}{82}\right)$ and a tf-idf value of $tf_{idf} = tf * idf = 0.41$

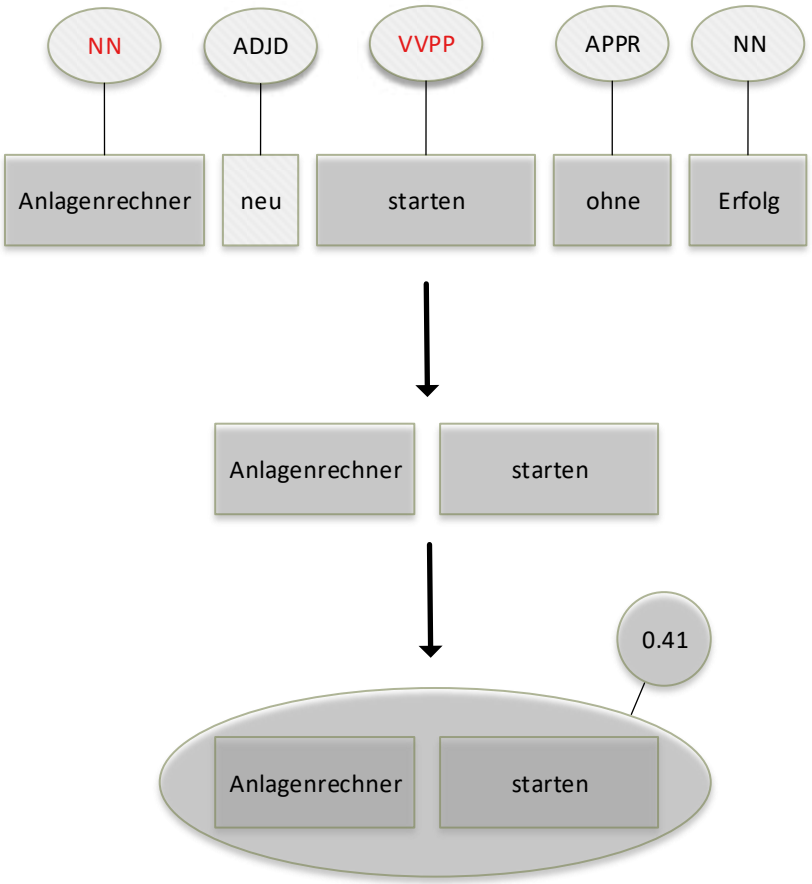

Figure 46: Model of keyword extraction and tf-idf calculation

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Applying the introduced concept of pre-processing, word embedding and keyword extraction results in a data table which is reduced in complexity and bulkiness as shown in Table 13. Notably, in report four the maintenance record was split in two sentences, also known as sentence splitting leading to two keyword phrases. However, in the provided data base the splitting of reports is rather scarce and plays a tangential role. The rule based extracted keywords from automatically generated logs were translated to German for gaining language consistency. Since English keyword expressions were in a manageable amount and small variety translation is achieved by dictionary approach.

Table 13: Extract of maintenance records represented by their keywords

EQUIPMENT	GUI_BUCHUNGSTEXT	DOWN_EVENT	AKTION	AUTHOR
L2301	zeitabweichung ätzen	ST	daten überprüfen	Person1
L2303	horden erkennung	ST	anlagenrechner starten	Person2
L2303	transport	ST	wafer entladen	Person3
L2302	zeitabweichung ätzen	ST	alarm auftreten wafer entladen	Person4
L2301	counter	ST	burn fahren	Person5

5.2.2 Implementation of Association Measuring

Association measuring pursues the goal of extracting knowledge from the pre-processed reports and investigates the dependencies between the extracted key words from different attributes. It is an approach for imitating the human ability of building associations based on the underlying text documents and is adapted from the introduced CIMAWA approach and Associative Gravity Force calculation in Section 2.2.3. In contrast to the provided data base, CIMAWA and AGF were tested and derived from comprehensive corpora composed of full and grammatically correct sentences but completely unstructured textual data. Thus, the measuring of the co-occurrence which is an essential part in the CIMAWA approach is based on a defined window size, investigating the keyword occurrence with respect to their relative positioning in text. As already mentioned in the introduction and illustrated in Table 12, the underlying data of this use-case is semi-structured by predefined meta-data represented in the columns of the table and by the provision of time stamps assigned to each row and report. Thus, the measuring of the co-occurrence is time and report-based, which means only keywords extracted from the same maintenance report and thus occurring at the same time are regarded for measurement. Since each reported maintenance operation corresponds to a new row in the data base, co-occurrence measuring is conducted row by row.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Another essential factor of CIMAWA and AGF is the tf-idf value measuring the importance of the keywords. Unlike CIMAWA, in the first place, keyword extraction is accomplished by part-of-speech tags forming noun and verb phrases instead of exclusively focusing on the biggest tf-idf values. The advantage of that approach is the involvement of morphological information of word types and sentence structure.[231] Especially in domain-specific texts the evaluation of word importance can be underestimated due to the vast usage of technical terms. Moreover, a body of reports drafted by various authors shows a higher probability for word repetitions than coherent texts. For instance, in the provided data the term “Wafer”, which in most cases belongs to the subject of the report, is frequently used resulting in a low tf-idf value. Omitting of that contextual important word could lead to a loss of information exacerbating the interpretation of the results by domain experts. Representing reports by a combination of noun and verb is a middle course of retaining the information content on the one hand but restraining the variety of key phrases on the other. Moreover, calculating tf-idf based on phrases -which is necessary for AGF measuring- overcomes the difficulty of importance underestimation since it increases the number of instances.

Another challenge of association measuring and co-occurrence measuring comes along with the German language and the linguistic possibility of compounding words. Using compound words adds information to a keyword sequence but might have a diminishing effect on the occurring frequency and could produce unique phrases. In the provided data the individuality of reporting comes along with the vast use of compounded words and different levels of detail in reporting. For overcoming that obstacle in the proposed procedural model frequency measuring is conducted along with word accordance checks. Regarding the following phrases “reinigungscounter erreichen” and “counter erreichen”, conventional frequency measuring would distinguish them as unique instances. Applying the accordance check enables recognizing the overlap of words which means “counter” is detected as part of the word “reinigungscounter”. Thus, the frequency measuring of “counter erreichen” also counts the compounded phrase “reinigungscounter erreichen”. However, regarding the opposite direction and investigating the frequency or co-occurrence of the compounded phrase, the implemented check separates them as unique instances. In other words, the compound term is modelled as a subset of the general word as shown below.

[231] cf. Khoury/Karray/Kamel, 2008, p.2

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

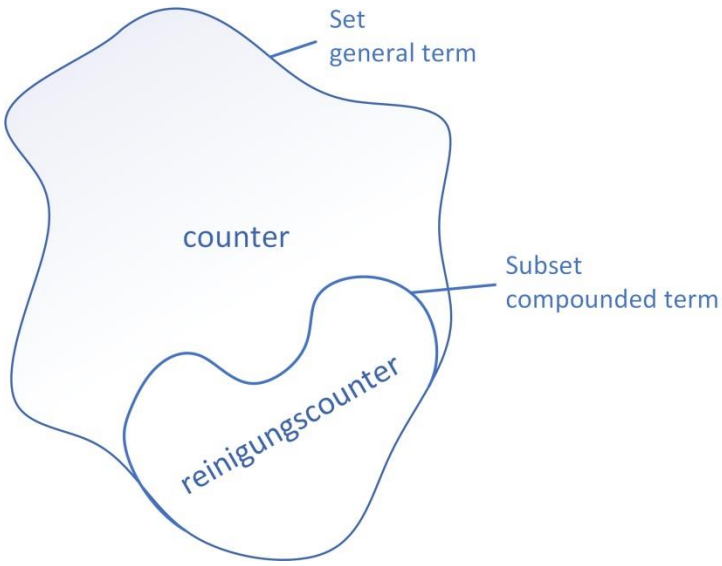

Figure 47: Co-occurrence measuring of compounded terms

The final step for accomplishing a satisfying frequency analysis is the correction of spelling mistakes. This model relies on the Levenshtein distance quantifying the difference between word strings. It is based on an algorithm computing the necessary edits of insertions, deletions and substitutions for transforming one word to the other.[232] In this application the Levenshtein distance is divided by the amount of characters in the strings and compared to a threshold. If the ratio falls below this value the sequences are regarded as equal. Thus, the longer the word the higher is the permissible deviation. If the character length is unequal, the smaller length is used in the comparison. A threshold of 0.15 has shown to provide suitable spelling correction.

$$\frac{Lev(a,b)}{\min\,(len(a), len(b))} < 0.15$$

Equation 33: Implemented Levenshtein ratio inequality

The mentioned adaptions do not affect the formulas of CIMAWA in its basic structure, but in this application the variables $word_1$ and $word_2$ also refer to key phrases. According to a comprehensive case study of Klahold et al. a dampingfactor in the interval of [0.4-0.6] achieves the best results. Thus, in this approach the dampingfactor is set to 0.5.

$$CIMAWA_{ws}^{\zeta}(x(y)) = \frac{Cooc_{ws}(x,y)}{(frequency(y))^{\alpha}} + \zeta * \frac{Cooc_{ws}(x,y)}{(frequency(x))^{\alpha}}$$

$$AGF(word_1(word_2)) = \frac{CIMAWA_{ws}^{\zeta}(word_1(word_2)) * kr(word_1)}{kr(word_2)}$$ [233]

Equation 34: Concept of CIMAWA and AGF

[232] cf. Gooskens/Heeringa, 2004, p.33
[233] cf. Klahold/Uhr/Ansari/Fathi, 2014, p.43

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

With the concept of associative gravity force the attraction between key words and phrases of maintenance report's attributes can be expressed. In the example below the Equipment "L2301" and the action "Anlagenrechner starten" are investigated. The thickness of the arrows is direct proportional to the attraction value, the diameter represents the importance. This example shows clearly the asymmetry of attraction which can be ascribed to the asymmetry of association. The visualization has to be read from arrow shaft to arrowhead. Starting from equipment "L2301" the attraction is rather weak whereas in the opposite direction the AGF achieves a higher value. From all conducted actions at equipment "L2301", "Anlagenrechner starten" seems to play a tangential role, but regarding all actions of "Anlagenrechner starten" leads to a strong interaction with "L2301". Having background knowledge about the actions and processes, maintenance operators are able to evaluate and interpret this additional information for deriving further conclusions. The tf-idf value contributes as a factor in AGF calculation and marks key words according to their importance being useful for detecting weak spots in reporting which will be dealt with in following sections. Finally, the concept of association measuring can be extended to all existing categories resulting in a network of keywords and phrases, as illustrated and commented in Figure 54 of Section 6.1.

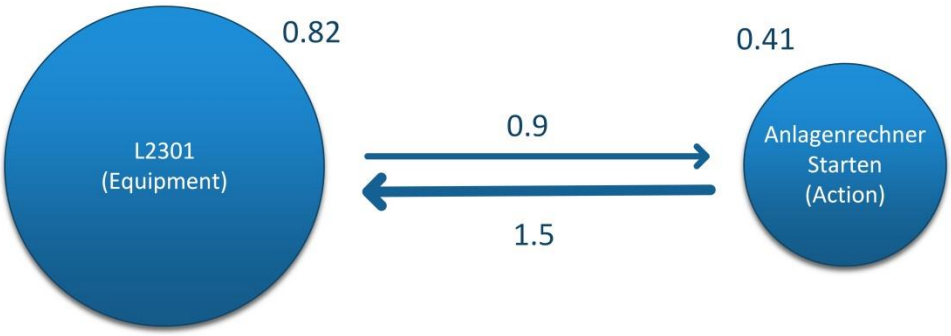

Figure 48: Schematic visualization of asymmetric AGF

5.2.3 Implementation of Opinion-mining and Sentiment Detection

The introduced approach of AGF is now extended by classifying the sentiments of the maintenance reports serving as an indicator for the previously extracted keywords. Since sentiment detection requires opinionated documents, the concept is only applied to texts of the column "Aktion". Nevertheless, subsequently the co-occurrences with the determined sentiment of corresponding attributes, such as equipment or author, can be investigated. The main idea behind this approach is to extract the background knowledge of maintenance staff gained by years and years of work, education and training. This kind of knowledge is implicitly hidden in the maintenance reports being expressed for example by appraisals, recommendations or opinions. Computational categorization of these opinionated documents according

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

to their sentiment is a way to explicate that knowledge and to label the corresponding key phrases based on their sentiment distribution.

This procedural model is based on a dictionary approach comparing the tokens to a set of opinion words classified in a positive and negative category. The lexical source was created by the University of Leipzig consisting of about 1650 positive and 1800 negative German opinion words.[234] The Sentiment detection is then conducted on the report level based on the classified tokens. However, inspection of the provided maintenance reports shows that the occurrence of typical opinion words such as “amazing” or “weak” is scarce and that the expressed sentiment is strongly linked to the domain-specific context. Regarding the following example report, neither positive nor negative opinion words can be found by using standard lexica, which leads to a neutral classification. However, the last word “erforderlich” or in English “required”, indicates the demand of additional actions and implies that troubleshooting has not been successful. Therefore, this report represents a negative sentiment in that individual domain but is considered to be neutral in a general context.

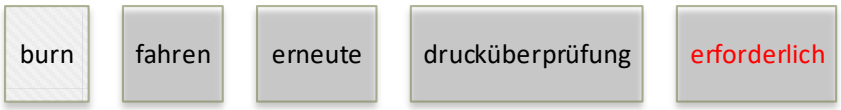

Figure 49: Example of negative opinion word

To cope with that challenge a domain specific dictionary is created considering special opinion words appearing in the application field of maintenance reporting. At first some of these words have to be manually selected. This includes words which express further requirements as the mentioned “erforderlich” but also terms which indicate requests such as “bitte” or outstanding actions like “ausständig”. Moreover, assessments such as “too high” are indicators of negative sentiments. The created set is then automatically grown by adding synonyms of the opinions words found in a synonym dictionary; this use-case draws on “Open Thesaurus”.[235] Together with the standard opinion lexicon the resulting list can finally be used for detecting negative opinion words in the reports. Analogously the procedure could be conducted deriving a positive opinion list. However, since the underlying data focusses on maintenance actions and corresponding errors, positive opinion words occur in a negligible amount. Thus, sentient detection becomes a binary model distinguishing between a neutral and a negative label.

[234] cf. Goldhahn/Eckart/Quasthoff, 2012
[235] cf. URL: https://www.openthesaurus.de, 04.09.2019

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Apart from terms in the mentioned opinion lists the word “ok” at the end of reports is regarded as a special indicator.

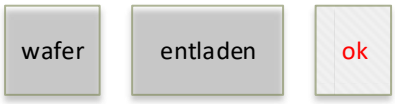

Figure 50: Maintenance action assessment

It implies that either the occurred error was, actually, a false alert or that the conducted action could finally fix the problem. Nevertheless, successful operations and troubleshooting is regarded as standard procedure and not always explicitly annotated. That is another reason why the distinction between a positive and a neutral category is refrained from and “ok” is not modelled as a typical positive opinion word. It is used for excluding a negative sentiment building a property in a set of rules for sentiment classification. In other words, the term “ok” at the end ensures the classification to the neutral category, even if there are preceding negative opinion words. All possible branches and their classification are listed in the following table:

Table 14: Classification branches of opinion words

negative opinion word	"ok" at the end of the report	category
		neutral
	x	neutral
x	x	neutral
x		negative

Although sentiment analysis is only applied to reports of the column “Aktion”, the sentiment co-occurrences of the corresponding attributes can be measured. This leads to a distribution for each instance or key phrase of negative and neutral classifications. For example, equipment L2301 is found in 1802 reports from which 1403 have a neutral and 399 a negative label.

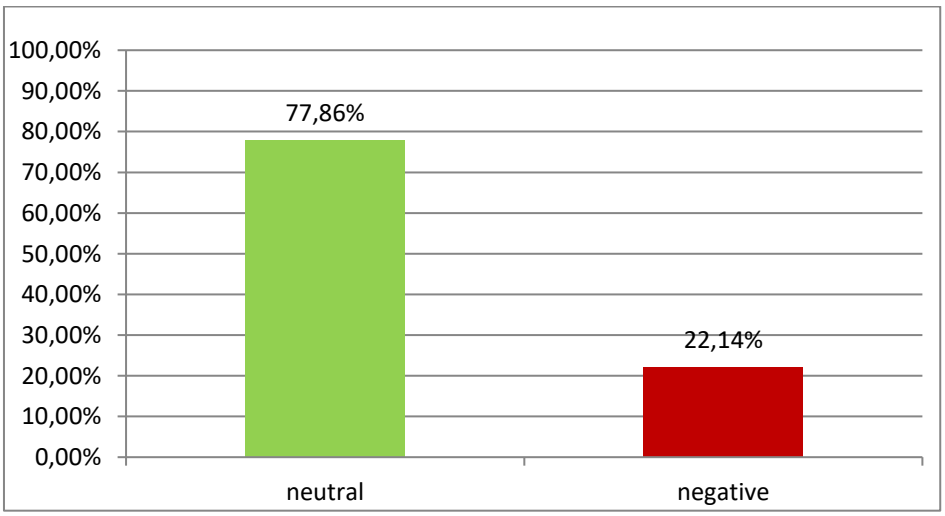

Figure 51: Sentiment distribution of equipment L2301

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.

This distribution can now be depicted by different colour shades from green representing 100% neutral classification to red illustrating pure negative labelling.

Figure 52: RGB based coloring for equipment evaluation

This method of visualization can now be easily combined with the previously mentioned concept of AGF revealing not only dependencies and the strength of attraction between key words, but also involving information about the maintenance staff's opinions and appraisals. This kind of domain knowledge is extracted across all report authors enabling the computational consolidation of various attitudes and opinions. Together with the extracted dependencies of AGF this evaluation can serve as a foundation for maintenance management in decision making. The results and comprehensive networks derived from the data are shown in Section 6.1.

5.3 Model Evaluation

According to the guidelines of artifact and model development of Alan R. Hevner evaluation and assessment can be achieved by the definition of appropriate metrics as well as the gathering of relevant data for measuring predefined attributes.[236] In this use-case the evaluation of the predicted sentiment classification is accomplished by the introduced Matthews correlation coefficient as well as Cohen's kappa based on the error types of the confusion matrix. Since there are about 10,000 binary classified reports, evaluation utilizes extrapolation for inferring from a sample to the population. The required sample size is estimated based on a confidence interval with a significance level of $\alpha = 5\%$ and the corresponding z-score of $z = 1.96$. The proportion $\hat{p}$ is unknown yet and estimated using an extract of 20 randomly selected reports. Three of them are assigned to a wrong category leading to a rough estimation of the proportion $\hat{p} = 0.15$. The required half width is defined as $h = 5\%$. Finally, the following equation determines the minimal sample size for the required boundaries. Based on that, the sample size is then set to $n = 200$.

$$n = \frac{z^2 * \hat{p} * (1 - \hat{p})}{h^2} = \frac{1.96^2 * 0.15 * (1 - 0.15)}{0.05^2} = 195.9$$

Equation 35: Sample size calculation for model evaluation

[236] cf. Hevner/Salvatore/Ram/Park, 2004, p.82

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

The randomly selected reports are manually inspected and assessed according to the error types of the confusion matrix distinguishing true positive, false positive, false negative and true negative classes. The confusion matrix of the considered sample is built as follows:

Table 15: Confusion matrix of regarded evaluation sample

Total population	Condition positive	Condition negative
Predicted condition positive	TP 136	FP 6
Predicted condition negative	FN 10	TN 48

The obtained classification can be used for evaluation through the Mathew's correlation coefficient and Cohen's Kappa. In contrast to the observed accuracy p_O these methods are statistical measures which also consider coincidental accordance. With a value exceeding the 80% level, both methods indicate satisfying classification accuracy. Thus, MCC expresses a high correlation between the predicted and the actual sentiment labels, whereas Cohen's kappa reveals high agreement between the classification set and the assessment corrected with the expected agreement by chance p_e.

$$p_O = \frac{TP + TN}{TP + FP + FN + TN} = 0.92$$

Equation 36: Observed accuracy of applied sentiment analysis

$$MCC = \frac{TN * TP - FN * FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} = 0.8027$$

Equation 37: MCC value of sentiment analysis evaluation

$$p_e = \frac{(TN + FP) * (TN + FN) + (FN + TP) * (FP + TP)}{(TP + TN + FP + FN)^2} = 0.5966$$

Equation 38: Expected accuracy of applied sentiment analysis

$$\kappa = \frac{p_O - p_e}{1 - p_e} = 0.8017$$

Equation 39: Cohen's Kappa of sentiment analysis evaluation

The applied evaluation technique belongs- according to the in Table 10 introduced methods- to the group of structural performance testing also referred to as white box testing. It considers internal mechanisms and requires knowledge about the underlying methods and algorithms. With respect to the applied evaluation of

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

sentiment analysis, the classification results of individual reports are compared, instead of evaluating the final outcome of sentiment distribution and corresponding coloring. This prerequisites knowledge about the implemented concept and about the predefined opinion words in order to detect program faults.[237]

The contrary approach of functional or black box testing exclusively evaluates the actual outcome and the visualized results of the software interface.[238] Applying black box testing to the output of the implemented model enables the assessment of derived associations and AGFs as well as annotated and depicted sentiments of key phrases. However, black box testing is interlinked with the business problem environment and requires experts familiar to the domain in order to accomplish evaluation. That part of evaluation is thus left open for the time being and can be finalized in cooperation with maintenance staff and management.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[237] cf. Gao/Tsao/Wu, 2003, p.141
[238] cf. Ibid., p.141

6 Summary and Consolidation of Key Findings

6.1 Results of Applied Methods

Applying the introduced approaches to the whole set of maintenance reports structured in Equipment, GUI, Aktion and Author and visualizing the outcome with respect to their attraction and sentiment, results in a network in which each node represents an instance of an attribute in the data base. Due to its bulkiness, the interpretation and visualization are limited to snap shots of the entire network. However, general rules and association patterns can be derived which are summarized in Table 16. At first, the results of association measuring and AGF are visualized and interpreted without involving sentiment detection, which should emphasize the difference to the extended model. Then, in the next step, the same keywords and phrases are regarded by including the labelling expressed by the color shades.

In Figure 53, a network of selected key phrases with all possible asymmetrical connections amongst them is depicted. The diameter of the nodes, of which each belongs to a different attribute, are directly proportional to the tf-idf numbers, the arrows' thickness expresses the strength of attraction. At first sight it is striking that the down event "ST" is represented by a much smaller circle than the other nodes implying low word weighting and importance. It occurs in a large number of documents and carries only little information, which results in a low tf-idf value annotating it as an unimportant word. This can be compared with articles in a text document occurring in a high amount but playing a tangential role in terms of text comprehension and information content. Another aspect which emphasizes this finding is the extremely asymmetric attraction towards "ST". This means, starting from "L2301", "Transport", "Person1" or "Wafer entladen" association measuring reveals a very strong dependency; however, the attraction is obviously weak in the opposite direction. The node "ST" is an important player from the perspective of the illustrated key words, but since the down event occurs together with a broad variety of other key words, one single key word is only attracted very weakly from the perspective of "ST". What can be learned from that circumstance is that there is a knowledge gap caused by insufficient reporting structure. The down event "ST" does not significantly contribute to knowledge representation, which could be improved by providing a more detailed classification of down events or subcategories. Thus, AGF visualizations provided to maintenance managers can reveal knowledge gaps of reporting structure and encourage them to take measures for improving maintenance documentation. With other words, the detection of those weak spots can be seen as targeted recommendations for reporting improvement.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Regarding the AGF between “Transport” and “Wafer entladen”, showing the interaction of the transport error and the discharging action for fixing it, the attraction is weaker than expected considering the logical dependence between them. Although the transport error and discharging action show a high co-occurrence, both instances are very frequently reported and are combined with multiple actions and errors. As a result, the regarded interaction is considered rather small in relative figures, which alleviates the attraction between them. The expectations of domain experts may, thus, not accord with the visualized results and reveal the range of usage of error terms serving as trigger for changes in reporting such as higher differentiation in error descriptions.

The attractions concerning the author “Person 1” strongly vary depending on the considered node interaction. Neither “Transport” nor “Wafer entladen” are remarkably attracted by the maintenance worker, which could be the case if workers are in charge of specific operations due to individual expertise in that domain. The attraction of the author and equipment appears stronger which might again be attributable to equipment expertise but could also originate from organizational issues such as shift schedules and areas of responsibility. In this case further interpretation and inferencing is left open for maintenance management possessing the required background knowledge.

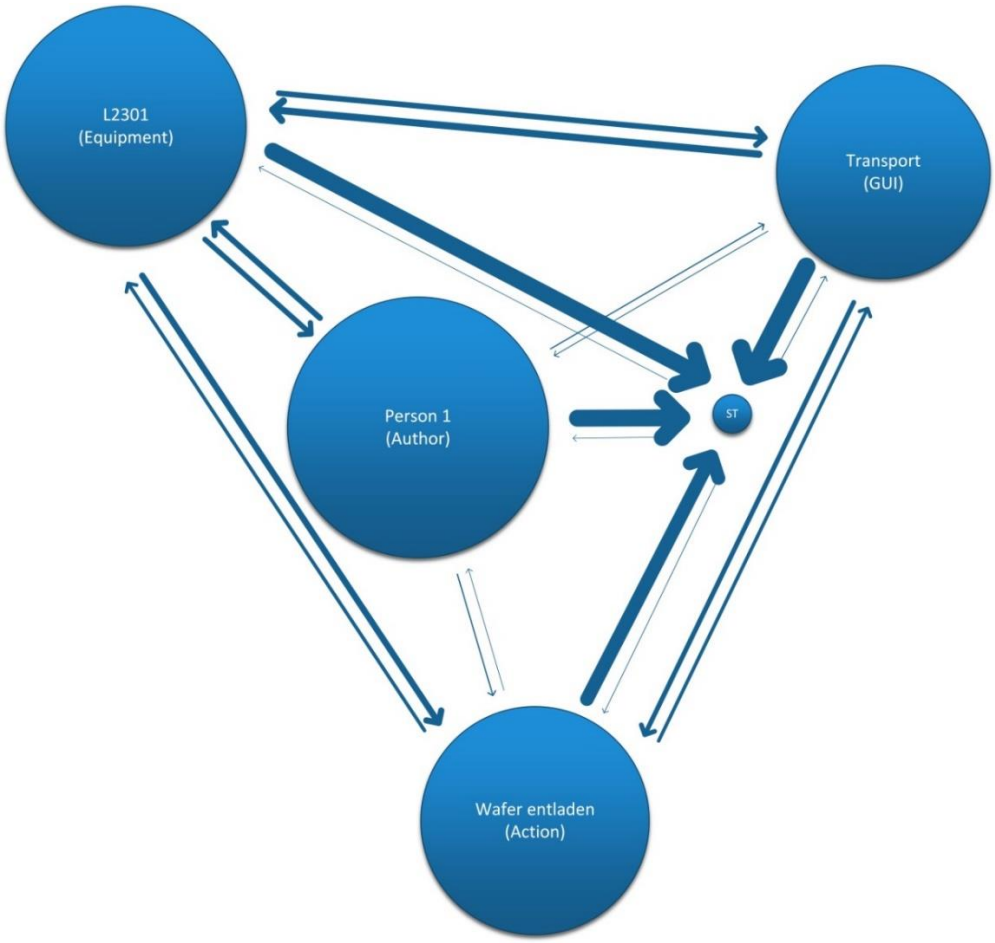

Figure 53: AGF network of extracted key phrases

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.

In Figure 54, the same network of AGF is extended by considering the sentiment of reports with respect to the individual key phrases. The generated labels depicted by the color shades from green to red express the distribution of neutral and negative classifications. From a general point of view, the figure is dominated by green coloring indicating a high amount of neutral classifications. Especially “Transport” and “Wafer entladen” are labelled by a lush green. The absence of negative sentiments marks them as uncritical incidents according to the appraisals and opinions of maintenance workers. One possible interpretation for that is that this kind of failures cause fewer inconveniences and can be fixed within a single maintenance operation. Thus, a strong force effect between equipment and transport errors can be seen with less concern. The involvement of all maintenance workers’ sentiments also provides an objective evaluation overview, which in environments of multiple plants, workers and operations is not always easy to keep track of. For example, the coloring of the equipment can be seen as an indicator of equipment condition regarded within certain time interval. Thus, also the labelling itself without the application of AGF already provides a general but useful evaluation.

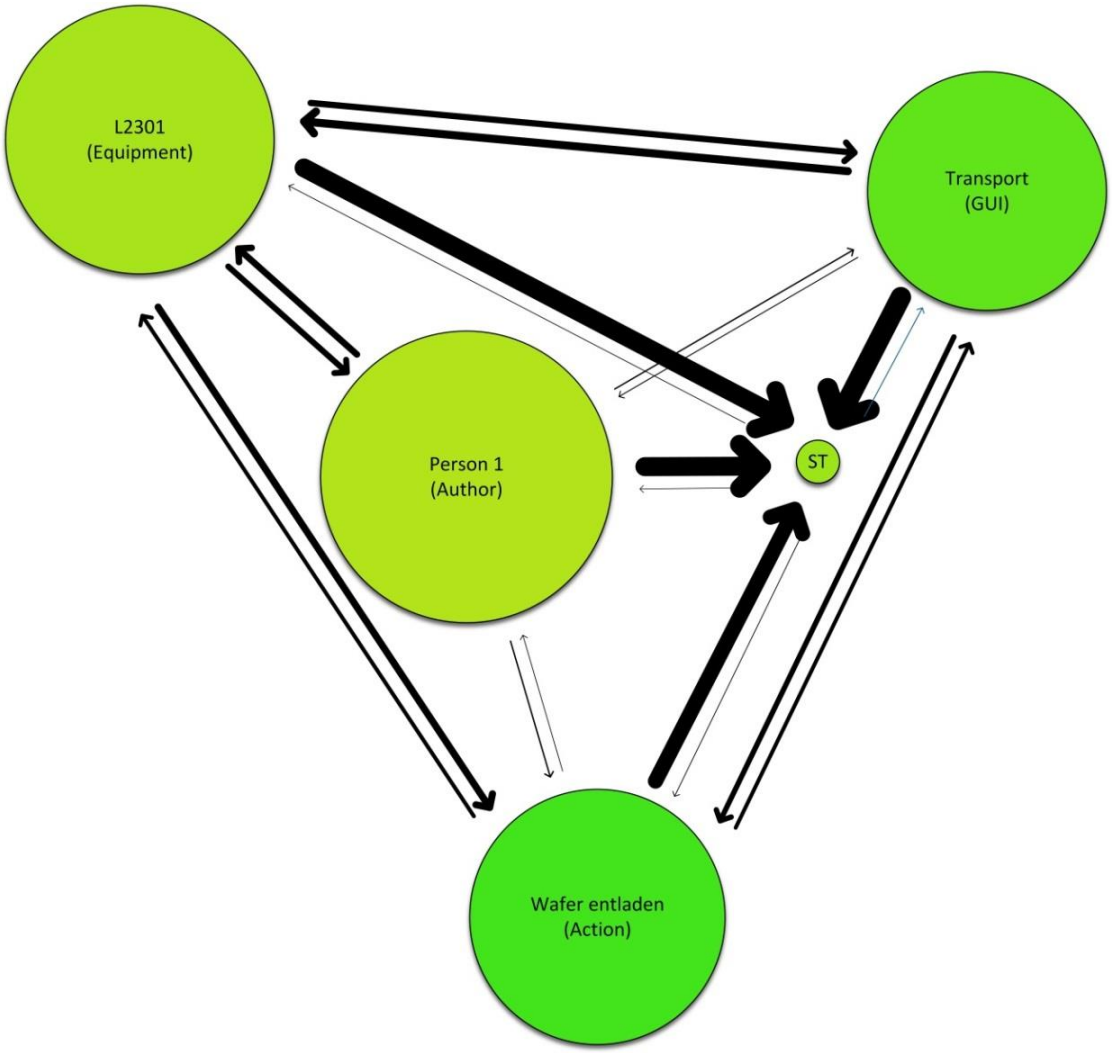

Figure 54: AGF and sentiment network of extracted key words and phrases

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.

For clarifying the concept of AGF in combination with sentiment labelling another network is illustrated which shows more differences in sentiment classification. The keyword extracted from the GUI column “Druck” or pressure is characterized by its orange label expressing a balanced distribution of neutral and negative classifications of pressure errors. The action “Gasspülen” comes along with a strong negative sentiment strikingly annotated by the red labelling. Compared to the previously investigated interaction of “L2301” and “Transport” in Figure 54, the AGF between “L2301” and “Druck” in this illustration shows similar relations. However, the additional sentiment information suggests paying particular attention to pressure failures and provides prioritizations between errors besides their associative gravity force. The same interpretation is applicable for “Gasspülen”. The strong attraction between equipment and action starting from “L2301” together with the annotation as critical event based on the opinions of operators encourages special treatment and consideration in management’s maintenance planning of that event. Regarding the opposite direction of AGF, originating from the action “Gasspülen”, the attraction appears weak indicating that the action is conducted at other equipment too.

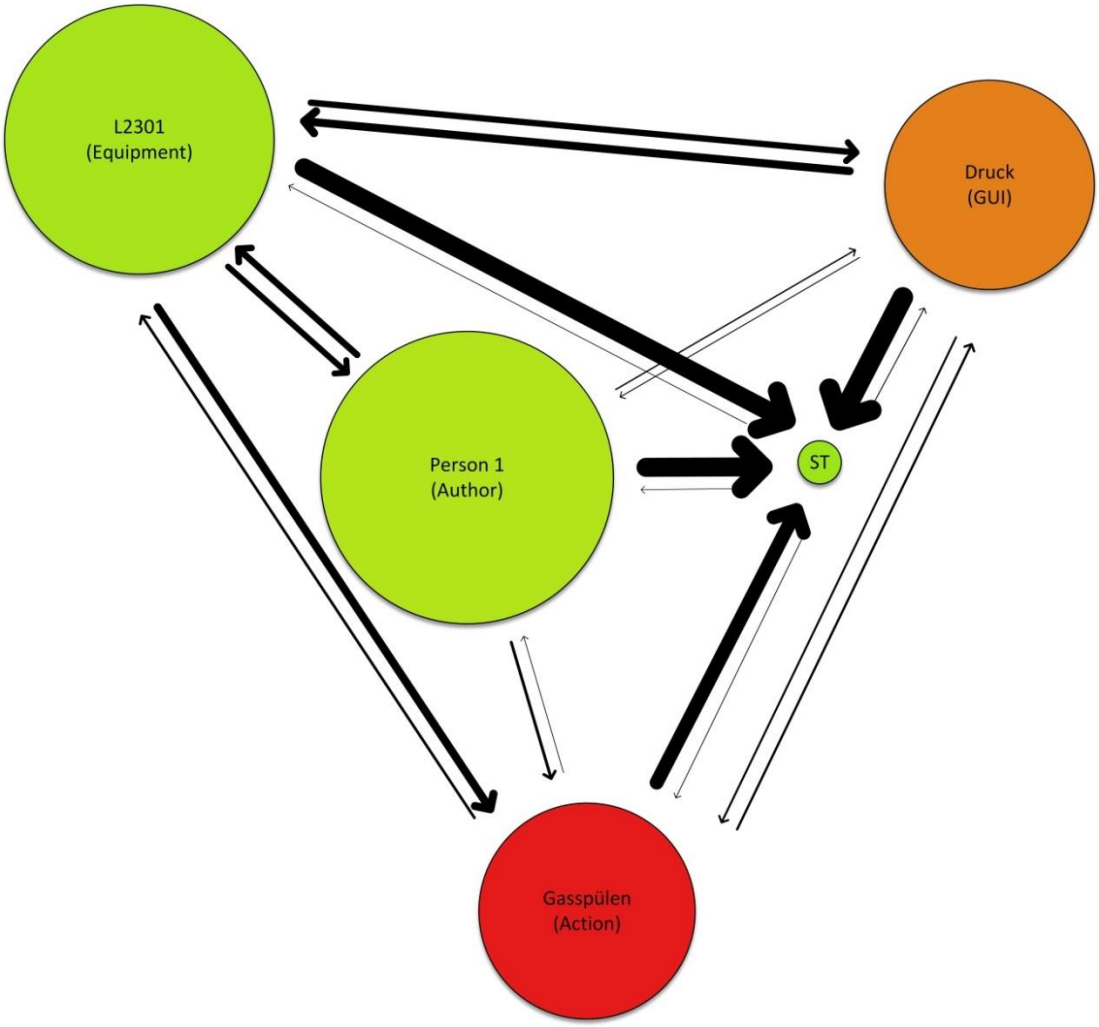

Figure 55: AGF and sentiment detection with various labels

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Of course, the net architecture of key words and phrases can be built up in various ways and does not necessarily have to be structured according to the report's meta-data. Figure 56 focusses on failures and their interdependency with equipment "L2301". Based on the AGF the transport error and "Ätzrate überschreiten", respectively etching rate exceedance, show the highest interaction from the equipment's perspective. Considering the sentiment labelling the etching rate error appears as the most critical error being represented by far more negative opinions compared to transport problems. Contrarily, the diameter of its circle expressing the importance appears rather small and the attraction is unsymmetrical. That aspects hint at a frequent occurrence across several plants instead of a specific "L2301" problem. Nodes such as "Spannung" and "Counter erreichen" show weak interactions but should not be ignored due to their labelling. All in all, that special network architecture reveals sensitivities towards failures of individual equipment based on AGF and operator's sentiment in maintenance reports. Providing these sensitivities to experts can encourage them to take appropriate measures for failure prevention and can help in building priorities for maintenance planning.

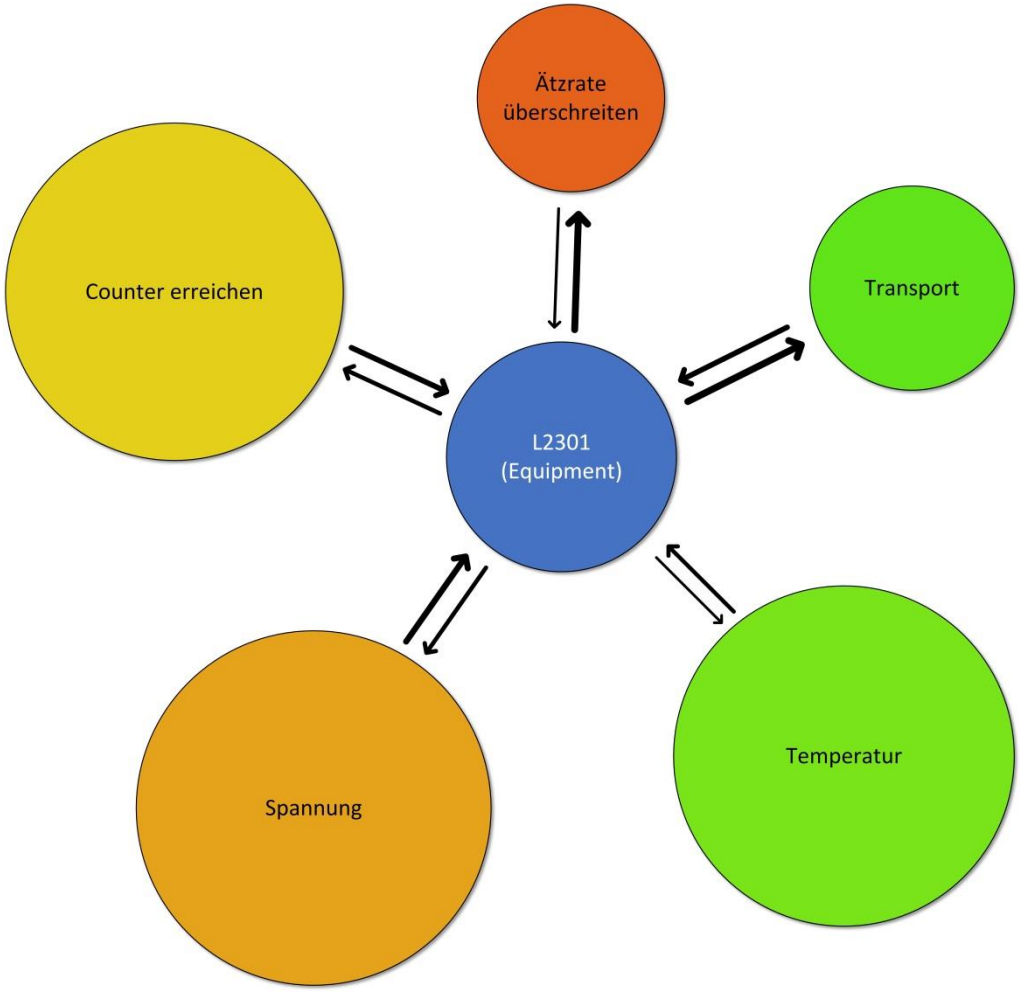

Figure 56: Failure analysis based on AGF and sentiment detection

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Table 16: Summary of detected rules and association patterns

Condition	Conclusion
Small node diameter	Low information content of key words
Strong asymmetric attraction	Weakly attracted node contains unspecific terms
Unexpected weak attraction	Key terms occur in multiple different events
Strong variation of attraction	Node depending interaction (e.g. different competences of employees/authors)
Node dominated by green coloring	Predominant classification to positive/neutral category
Node dominated by red coloring	Predominant classification to negative category

6.2 Results with Respect to Research Questions

This section refers to the research questions stated in Chapter 1.2 and provides comprehensive answers for each of them.

Q1: Does the integration of text-mining significantly contribute to improvement and enrichment of prescriptive maintenance models?

Currently, predictive and prescriptive maintenance approaches mainly focus on real-time data acquired by condition monitoring systems using sensors for detecting machinery anomalies.[239] In other words, the main focus is on the creation of new knowledge and on finding coherences from object based data by applying advanced data analysis methods.

Conversely, the investigated approach of text-mining concentrates on a knowledge component which is rather neglected in the area of industrial maintenance so far; existing and human-related knowledge. Text is one of the main mediums of human communication and thus captures ideas, thoughts, experiences etc. In the environment of industrial maintenance text implicitly contains background knowledge of maintenance experts which is expressed by appraisals, recommendations and opinions but also coherences and causalities of representative key words. Explicating this kind of knowledge reveals insights which are not retrievable by solely considering object and machinery related data. Thus, the integration of text-mining to prescriptive maintenance enables access to valuable human background knowledge enriching the existing object-data-driven approaches. On a larger scale, text-mining models in industrial maintenance as introduced in Section 5, can build the missing link between equipment and humans combining the existing and the generated knowledge, which is visualized in Figure 57.

Since mismanagement in maintenance is one of the root causes of arising troubles being tied to a lack of comprehensive and mature knowledge[240], knowledge

[239] cf. Gopalakrishnan/Kar/Bose/Basu, 2019, p.1
[240] cf. Ansari/Glawar/Nemeth, 2019, p.1

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

discovery from text can be utilized for improving management activities and supporting decision making. Thus, the provision and visualization of a comprehensive knowledge base to maintenance managers has positive effects on performance and productivity, which means the application and integration of appropriate text-mining models also achieve improvements from a profitability point of view.

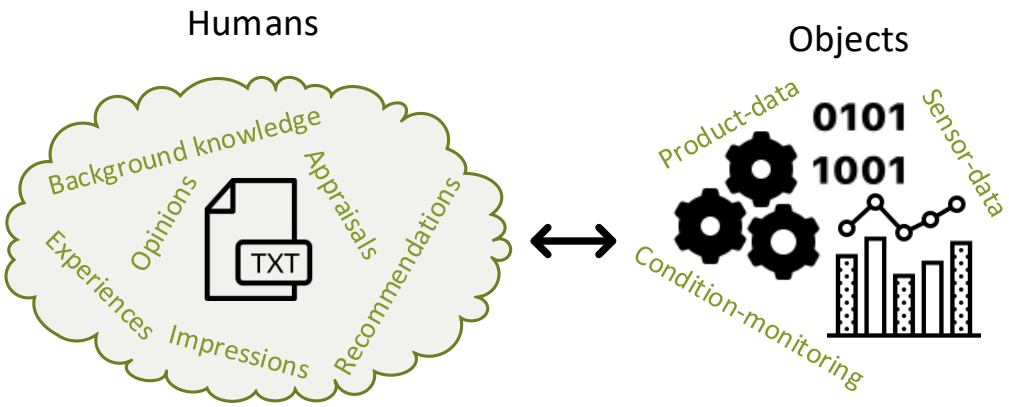

Figure 57: Linkage of human and object related knowledge assets

Q2: Which methods succeed in explicating implicit knowledge and detecting hidden patterns in industrial maintenance?

Since text is an unstructured data source characterized by individual style and preferences of the author as well as word variations based on grammatical rules, explication of implicit knowledge first requires pre-processing methods in order to prepare the textual data for analysis. Pre-processing methods include techniques for splitting sentences into tokens, removing unnecessary punctuation and stop words, finding root forms of words and lowercase obtained tokens. Depending on the following methods of keyword extraction and analysis, pre-processing also fulfills the task of adding morphological information to tokens, such as part-of-speech, language and tense. The utilized methods and algorithms for achieving text pre-processing, range from dictionary approaches to machine learning techniques such as neural networks and probabilistic approaches like hidden Markov chains.

Another set of methods necessary for knowledge extraction and for preparing the text data for the actual analysis are word embedding and word representation methods. There are two common concepts for converting text to numerical representation; Frequency-based methods such as Bag of Words or term-frequency inverse document-frequency which count the occurrence of words and weight them according to their global and local frequency. Word2Vec methods which convert each word to a vector by trying to predict target words based on their context words and building a vector space expressing word meanings and similarities.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Textual data in the area of industrial maintenance such as maintenance protocols and reports are valuable because of their latent knowledge implicitly hidden in them. Maintenance reports not only serve the purpose of documentation, but also express opinions and appraisals drafted by the maintenance staff. Sentiment analysis or opinion mining provides methods for mining and quantifying these opinionated text documents towards their sentiment and are thus an advisable technique for knowledge explication. Sentiment analysis is accomplished by the development of domain specific dictionaries providing positive and negative opinion words for classification. This ensures the consideration of contextual ambiguities and domain specific word sentiments. Supervised machine learning algorithms- especially convolutional neural networks- are another promising possibility for implementing sentiment detection. They stand out by their ability of learning structures and patterns from a labelled data set classifying text documents according to predefined features.

Apart from opinions and sentiments, implicit knowledge also exists in the form of latent inferences and causalities. Association measuring is a concept for deriving word inferences from text by imitating the human ability of building associations. It is based on the co-occurrence and absolute frequency of pre-extracted keywords and also considers the asymmetrical nature of word association. Combining the methods of word association and word importance measured by the concept of tf-idf, the attraction of keywords can be expressed, which is referred to as associative gravity force. This concept quantifies and explicates the hidden interaction of key words resulting in a net representing attractions as well as individual word importance. It is thus an appropriate method for deriving hidden patterns and explicating latent knowledge in the domain of industrial maintenance.

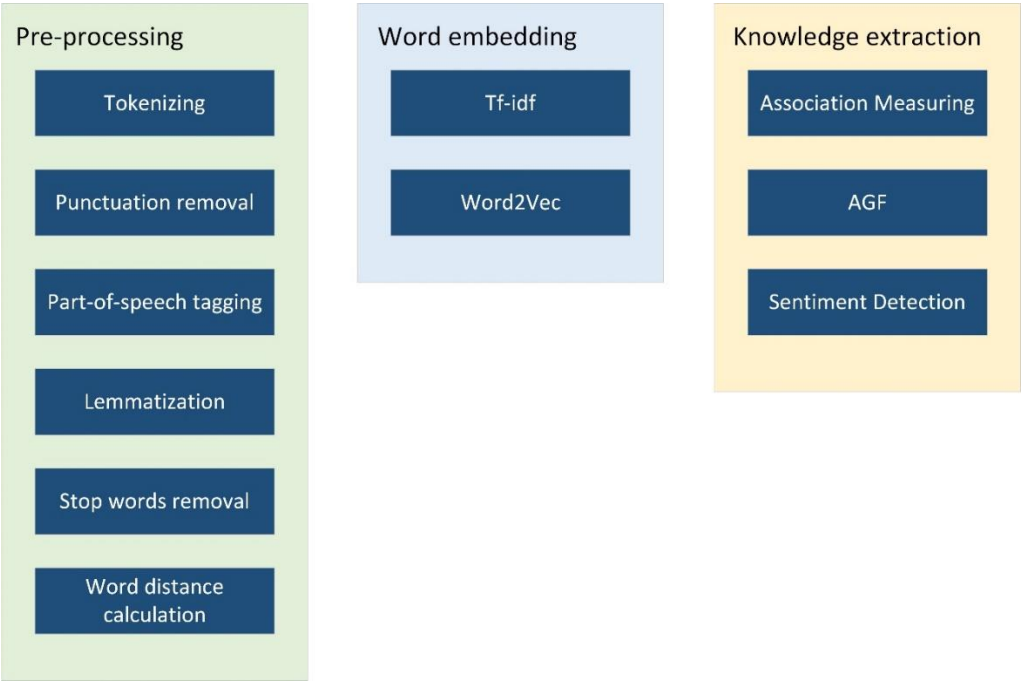

Figure 58: Text-mining methods structured according to purpose

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Q3: How should the results being effectively communicated and visualized?

There are three main instances of extracted information which have to be clearly and effectively communicated, as shown in Figure 59. The importance of key words based on the global and local weighting, the attraction between them, which in general appears asymmetrically, and the sentiment classification of the individual key word. It is desirable to gather all that information in a single visualization model in order to facilitate comprehension for decision makers. On the contrary, the observer and knowledge recipient should be able to distinguish between the different types of information. Beginning with the aspect of keyword attraction, due to its asymmetry and anisotropy arrows are appropriate symbols for visualization, whereas the shaft can express the direction of attraction and the thickness the strength of force. The keywords themselves are located in circles connected by these arrows. The diameter of the circle is proportional to the importance of the key word combining the aspects of attraction and word importance. Finally, the information of sentiment has to be added to the obtained AGF network. In order to preserve compactness, the visualization concept remains unchanged in its geometrical shape but is extended by including the dimension of coloring. This means, each circle is assigned a certain shade of color corresponding to the analyzed sentiment distribution, accordingly, red expresses a pure negative sentiment distribution and green a positive one. Thus, coloring orientates on common and intuition-based conventions, using the red signal color as expression for danger and negative aspects, whereas green is associated with correctness and positive sentiments. Interpolation between those states can be achieved by a linear model based on the RGB color codes. The proposed depiction uses a two-dimensional approach but can in case of more complicated networks and space problems be extended to the third dimension.

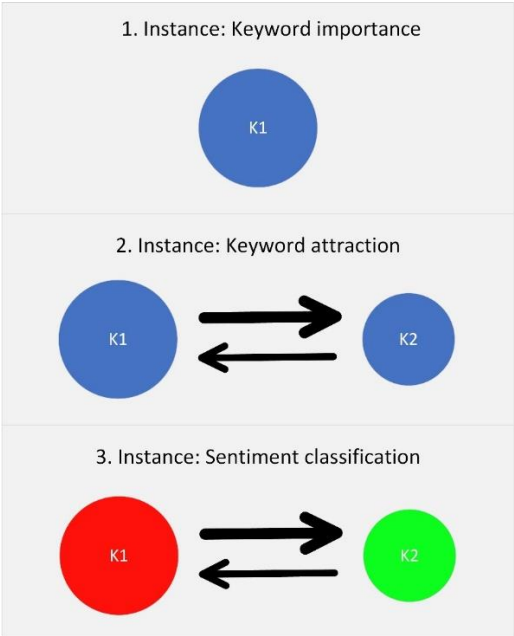

Figure 59: Visualized instances of information

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Q4: How to recommend standard reporting structure and template including contents for typical problems to improve quality of text reports?

Since reporting in industrial maintenance is mainly based on unstructured or semi-structured text documents, level of detail, comprehensiveness of reporting, style and way of expression as well as correctness vary depending on the individual author. As a consequence, knowledge extraction becomes a challenging task requiring numerous pre-processing steps in order to prepare the textual data for analysis. Prevalent methods for overcoming this obstacle are templates and standards for reporting. The proposed method of knowledge extraction can serve here as a foundation for template creation revealing the importance, interaction and sentiment towards extracted key words. First of all, the extracted key words can be utilized as predefined text modules listed, for example, in a drop-down menu. Each of them is assigned a tf-idf value showing the importance and level of information content. If the diameter of a key word is very small, it is advisable to either remove it or replace it by a more detailed set of categories since its tf-idf weight and, respectively, the information content is very low. Low tf-idf values often also come along with strong asymmetrical attraction being another indicator for possibilities of improvement in reporting structure. But also the sentiment of a key word can be decisive for the way of reporting. Red colored circles express critical instances such as actions or failures and should be considered more than uncritical standard procedures. Reporting can be adapted to the sentiment and urgency of keywords for example by providing a more detailed template for critical maintenance actions. Since knowledge extraction and its visualization is a dynamic process yielding different results over time, template creation and reporting has to be adapted to the changing environment. As illustrated in the overall procedural model architecture in Figure 38 template creation is up to the maintenance manager based on the explicated and visualized knowledge being embedded in a continuous circle of improvement and adaption.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

7 Outlook and Limitations

The introduced procedural model is designed for extracting knowledge from semi-structured maintenance reports by a broad variety of text-mining methods. Due to the specificity of used vocabulary and underling data, pre-processing tasks as well as sentiment analysis are accomplished by especially created dictionaries. Thus, applying the concept to textual maintenance data with a different kind of structure and vocabulary requires adaptions for successful implementation. A possibility which is left open for further discussion for improving standardization and general applicability is realizing knowledge representation by ontologies. The definition of an ontology is given as follows: “An ontology is a formal, explicit specification of a shared conceptualization”[241]. The term conceptualization refers to an abstract model representing a phenomenon in the word, formal means it is machine readable, specifications include properties, functions, relations, constraints etc. being explicitly defined.[242] Ontologies thus represent the basic terms and their relations of a topic area as well as the rules to extend this vocabulary.[243] Applied to industrial maintenance ontologies can represent functions and interconnections of objects, systems and subsystems such as hydraulics, electronics, and pneumatics but also hoses, pumps and fluids. This domain model provides natural categories which can be utilized for example by cluster algorithms to classify maintenance logs to their functional groupings.[244]

Numerical word representation of text in the proposed approach is achieved by the tf-idf weighting scheme expressing the importance of a word according to its frequency. Alternative approaches are word-embedding models such as word2vec converting each word to a vector representing its meaning in a vector space (see chapter 2.2.2). This advanced concept could be used for clustering the maintenance reports with respect to their message and generating maintenance categories. However, the currently provided reports were generated as free text input by maintenance staff without any support by templates and structural guidelines for drafting. As a consequence, reports strongly differ in level of detail and style which means records may even be composed of just a single word. Since word embedding methods build the vector representation based on the surrounding context words, an approximately identical level of detail in reporting is desirable. Consequently, the application of word embedding methods to the introduced record data leads to unsatisfying results. One approach for overcoming the obstacle of incomplete reporting is the communication of guidelines for drafting. It would be purposeful to provide information about the

[241] Studer/Benjamins/Fensel, 1998, p.25
[242] cf. Ibid., p.25
[243] cf. Bernaras/Laresgoiti/Correra, 1996, p.298
[244] cf. Ram/Devaney/Qiu/Lee, 2005, pp.1f

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

main attributes of maintenance operation, namely, equipment, error type, conducted action and final assessment. Achieving comprehensiveness of reporting could then enable further steps of analysis and model advancement by using the mentioned word embedding methods.

Using templates and prepopulated forms with contents of typical problems is a more rigorous way for implementing the previously mentioned guidelines for increasing standardization in maintenance reporting. Although there is an enormous effort for accomplishing pre-processing and cleansing of tokens in the proposed approach as demonstrated in Section 5.2.1, there is still a remaining amount of words, which contains spelling mistakes or could not be sufficiently prepared. Implementing templates and forms could reduce that pre-processing efforts whilst ensuring an adequate level of data quality. The conducted association measuring and sentiment analysis visualize the most important but also most critical key words which are interesting for prepopulating forms and templates. Based on that information the simplest way of creating a standard reporting form is by using drop-down lists for each operation attribute. One schematic realization showing an extract of maintenance actions and four reporting attributes is depicted in Figure 60. If errors and maintenance actions appear in a broad variety, drop down menus could be additionally divided in subcategories based on, for example, machinery components for increasing practicability and clearness. Moreover, it has to be pointed out that prepopulated key words have to be continuously checked and adapted by maintenance staff for obtaining effective reporting support. Designing drop down lists with an optional free text input field could avoid inappropriate selection in case of unforeseen events and give the opportunity to report yet unstandardized operations.

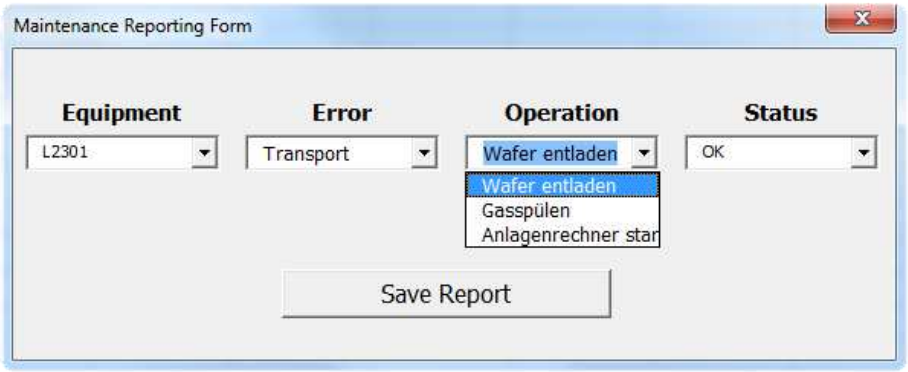

Figure 60: Possible example of a maintenance reporting form

Apart from the intensively discussed textual and numerical data, images could be utilized as additional source supporting the representation of extracted background knowledge. Currently one of the main purposes for taking pictures in industrial areas is the documentation of failures and critical events with the ulterior motive of liability. Images are then usually stored using document management systems but not further

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

exploited. Extending the reporting of maintenance activities and failures by taking pictures and applying the introduced concept to that enhanced data base, images could be assigned to the extracted key words based on their co-occurrence. In the resulting AGF network key words are then additionally described by corresponding images and are providing further information about the regarded action or failure. This helps to understand the meaning of the key word as well as to reveal problems arising through word ambiguities. Of course, model and data extension could also be achieved by exploiting other information sources such as audio data or film sequences. The methodology of association measuring and the basic idea of information annotation to extracted key words, however, remains the same.

Since one of the main goals of this work involves the enriching of sensor- and condition-based maintenance planning by mining textual sources, the results obtained from these different data perspectives or respectively the data itself have to be merged. In the proposed procedural model extracted knowledge is visualized and presented to maintenance mangers for decision support. The consolidation is thus realized on the level of maintenance management which is provided knowledge from different data perspectives. However, apart from receiving visualized results and making decisions the management level accomplishes the task of output-revision. The derived inferences from association measuring are based on CIMAWA relying on the co-occurrence of keywords. Consequently, detected dependencies might be subject to a spurious relationship and do not necessarily express causality. Maintenance staff and management possess the required background knowledge for assessing the acquired relations and can finally draw the correct conclusions for decision making. Visualization thus serves as an instrument of revision and is not only limited to the communication of results.

Combining textual data and numerical data at an earlier stage of input-vector generation for the processing by machine learning algorithms could be an additional approach for model advancement. Chakraborty and Pagolu state that the involvement of textual data improves the predictive accuracy of these numerical data models.[245] The quantified sentiment towards equipment and other entities as well as textual maintenance records converted by word embedding methods could be utilized for enriching numerical machine learning models in the area of industrial maintenance. The desired outcomes involve the previously mentioned detection of failure and measurement categories which is a typical problem of unsupervised machine learning as well as the prediction of the remaining lifetime of equipment components. On a larger scale, the improvement of accuracy should finally effect coordinated and well-planned maintenance scheduling and have an impact on overall manufacturing profitability.

[245] cf. Chakraborty/Pagolu, 2014, p.2

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

The visualized results of association measuring and sentiment analysis as depicted in Section 6.1 are presented in a two-dimensional AGF network with colored nodes. Focusing on one representative of each main category with a limited amount of interactions and dependencies, the chosen visualization method provides clear and comprehensible illustrations. However, investigating more complex structures by regarding multiple instances of each main category and their corresponding connections, the two-dimensional depiction might become confusing and even for domain experts difficult to comprehend. One attempt for coping with complex association structures is expanding the illustration to the third dimension and representing them by AGF sphere networks. The same principle of arrow scaling AGF expression and sentiment coloring is then applied to spheres providing a larger surface for dependency illustrations. Of course, the maintenance manager and observer have to be enabled to navigate through the network and regard the results from different perspectives and distances. Otherwise, three-dimensional visualization would rather worsen knowledge representation in terms of effectiveness and clearness.

In terms of application environment, the introduced approach could be of interest for digital reporting software such as the commonly used Finito Digital Shift Book from the German company New Solutions.[246] Currently the software provides comprehensive tools and interfaces for standardized and cross-factory reporting and documentation of shift events[247]; however, text-mining solutions as proposed and investigated in this work are not implemented yet. Integrating text analysis to software for standardized reporting is a promising concept since text-mining algorithms can be applied to documents of high data quality. The proposed procedural model can thus be utilized for advancing and extending existing documentation and reporting software to a tool of knowledge extraction and decision support in industrial maintenance.

A topic of recent research is the support of complex maintenance tasks using augmented reality.[248] In contrast to virtual reality (VR), which simulates an environment with complete separation from the actual reality, augmented reality recognizes the appearance of objects of the real world and enriches them by visualizing related computational information.[249] This kind of information can be, for example, the current condition of equipment but also textual instructions for maintenance actions. The approach of Engelke et al. even includes the creation and attachment of text notes to registered and 3D-modeled machinery enabling new possibilities in maintenance reporting.[250] Referring again to the presented use-case,

[246] cf. URL: https://www.schichtbuch.com, 18.11.2019
[247] cf. Ibid.
[248] cf. Engelke/Keil/Rojtberg/Wientapper/Schmitt/Bockholt, 2013, p.105
[249] cf. Dörner/Broll/Grimm/Jung, 2013, p.2
[250] cf. Engelke/Keil/Rojtberg/Wientapper/Schmitt/Bockholt, 2013, p.107

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

the results of conducted sentiment analysis towards equipment expressed by color shades could be annotated to currently observed machinery using the visualization method of AR. Moreover, the AGF connections of most occurring and most critical events could be depicted in immediate proximity of recognized equipment. The computational information can thus be overlapped with the objects of real-world replacing terminals and interfaces. The following figure shows an example of AR in maintenance annotating text and state conditions to machinery.

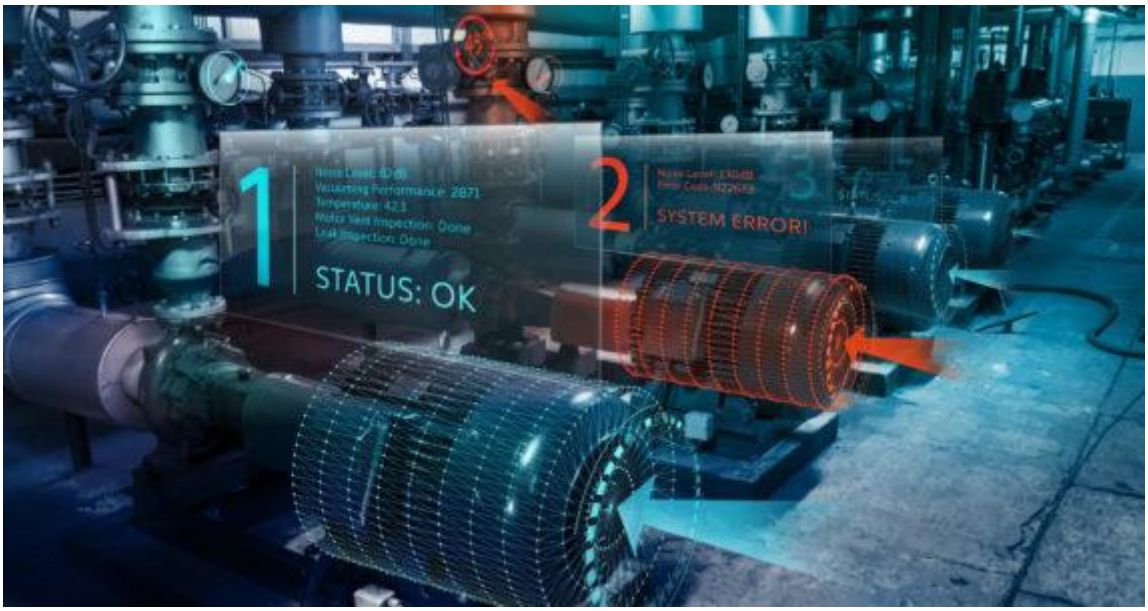

Figure 61: Augmented Reality application in industrial maintenance[251]

Finally, Table 17 summarizes proposed future work and mentioned research directions which could contribute to improve and enhance the presented procedural model.

Table 17: Summary of proposed research directions for model enhancement

Nr.	Proposed future research directions for model enhancement
1	Industrial maintenance knowledge representation by ontologies
2	Semantic word vector representation by word-embedding methods (e.g. Word2Vec)
3	Template and form creation for standardized maintenance reporting
4	Improvement of maintenance support by image annotation
5	AGF and keyword visualization by 3D sphere networks
6	Implementation and integration to existing reporting software (e.g. Digital Shift Book)
7	Visualization of results by augmented reality

[251] cf. URL: https://www.cbronline.com/in-depth/ar-goggles-for-enterprise, 22.10.2019

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

8 References

Abate H.: The PF Interval. URL: https://www.assetivity.com.au/article/reliability-improvement, 25.05.2019

Aggarwal C. C.; Zhai C.: Mining Text Data, Springer US, Boston, 2012

Aghdam H.; Heravi E.: Guide to Convolutional Neural Networks, Springer International Publishing, Cham, 2017

Alexandrov D. A.; Boukhanovsky A. V.; Chugunov A. V.; Kabanov Y.; Koltsova O.: Digital Transformation and Global Society, Springer International Publishing, Cham, 2018

Alharbi A. S.; Doncker E.: Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information, in: Cognitive Systems Research, Vol. 54, 2019

Anandarajan M.; Hill C.; Nolan T.: Practical Text Analytics, Springer International Publishing, Cham, 2019

Ansari F.: Cost-Based Text Understanding to Improve Maintenance Knowledge Intelligence in Manufacturing Enterprises, in: Computers & Industrial Engineering, 2020

Ansari F.; Glawar R., 2019: Knowledge-Based Maintenance, in: Instandhaltungslogistik, Qualität und Produktivität steigern, 7. edition, editor: Matyas K., Hanser, München, pp. 318–342

Ansari F.; Glawar R.; Nemeth T.: PriMa: a prescriptive maintenance model for cyber-physical production systems, in: International Journal of Computer Integrated Manufacturing, Vol. 7/1, 2019

Ansari F.; Uhr P.; Fathi M.: Textual meta-analysis of maintenance management's knowledge assets, in: Economics and Management, Vol. 6/1, 2014

Arif-Uz-Zaman K.; Cholette M. E.; Ma L.; Karim A.: Extracting failure time data from industrial maintenance records using text mining, in: Advanced Engineering Informatics, Vol. 33, 2017

Ben-David A.: Comparison of classification accuracy using Cohen's Weighted Kappa, in: Expert Systems with Applications, Vol. 34/2, 2008

Benesty J.; Sondhi M. M.; Huang Y.: Springer handbook of speech processing, Springer, Berlin, London, 2008

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Bernaras A.; Laresgoiti I.; Correra J.: Building and Reusing Ontologies for Electrical Network Applications, 12th European conference on Artificial Intelligence, Chichester, 1996

Biedermann H.: Predictive Maintenance, Realität und Vision : 32. Instandhaltungsforum, TÜV Media, Köln, 2018

Bird S.; Klein E.; Loper E.: Natural Language Processing with Python, O'Reilly Media, Sebastopol, 2009

Blanco E.; Dan M.: Some Issues on Detecting Negation from Text, 24th International Florida Artificial Intelligence Research Society Conference, Dallas, 2011

Buneman P.: Semistructured Data, University of Pennsylvania: Department of Computer and Information Science, 1997

Chakraborty G.; Pagolu M.: Analysis of Unstructured Data- Applications of Text Analytics and Sentiment Mining, Oklahoma State University: Department of Marketing, 2014

Chan F.T.S.; Lau H.C.W.; Ip R.W.L.; Chan H. K.; Kong S.: Implementation of total productive maintenance: A case study, in: International Journal of Production Economics, Vol. 95/1, 2005

Chapman P.; Clinton J.; Kerber R.; Kahabaza T.; Reinartz T.; Shearer C.; Wirth R.: CRISP-DM, Step-by-step data mining guide, 2000

Chen Y.; Wang W.; Liu Z.; Lin, Xuemin: Keyword Search on Structured and Semi-Structured Data, Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, 2009

Cho K.; van Merrienboer B.; Gulcehre C.; Bahdanau D.; Bougares F.; Schwenk H.; Bengio Y.: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, Université de Montréal: Institute for Data Valorization, 2014

Denham K.; Lobeck A.: Linguistics for everyone, An introduction, Wadsworth Cengage Learning, Boston, Mass., 2010

Deutsches Institut für Normung: Grundlagen der Instandhaltung, DIN/31051, 2018

Dörner R.; Broll W.; Grimm P. F.; Jung B.: Virtual und Augmented Reality, Grundlagen und Methoden der Virtuellen und Augmentierten Realität, Springer, Berlin, Heidelberg, 2013

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Dörre J.; Gerstl P.; Seiffert R.; Fayyad, Usama: Text mining: Finding Nuggets in Moutains of Textual Data, Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999

DZone Inc.: The Levenshtein algorithm. URL: https://dzone.com/articles/the-levenshtein-algorithm-1, 11.09.2019

Engelke T.; Keil J.; Rojtberg P.; Wientapper, Folker; Schmitt, Michael; Bockholt, Ulrich: Content First, A concept for Industrial Augmented Reality Maintenance Applications using Mobile Devices, International Symposium on Mixed and Augmented Reality, Adelaide, 2013

Ertel W.: Introduction to Artificial Intelligence, Springer International Publishing, Cham, 2017

Fawcett T.: An introduction to ROC analysis, in: Pattern Recognition Letters, Vol. 27/8, 2006

Fitch E. C.: Proactive Maintenance for Mechanical Systems, Elsevier Science, Burlington, 1992

Flasiński M.: Introduction to Artificial Intelligence, Springer International Publishing, Cham, 2016

Frangopol D. M.; Lin K.-Y.; Estes A. C.: Life-Cycle Cost Design of Deteriorating Structures, in: Journal of structural engineering, Vol. 10/123, 1997

Furbach U.; Krämer T.; Schon C., 2019: Names Are Not Just Sound and Smoke, Word Embeddings for Axiom Selection, in: 27th International Conference on Automated Deduction, editor: Fontaine P., Springer, Cham, Switzerland, pp. 250–268

Gao J.; Tsao H.-S. J.; Wu Y.: Testing and quality assurance for component-based software, Artech House, Boston, 2003

Gao R.; Wang L.; Teti R.; Dornfeld D.; Kumara S.; Mori M.; Helu M.: Cloud-enabled prognosis for manufacturing, in: CIRP Annals, Vol. 64/2, 2015

Goldhahn D.; Eckart T.; Quasthoff U.: Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages, 8th International Language Ressources and Evaluation conference, 2012

Google LLC: Language detection. URL: https://code.google.com/archive/p/language-detection/, 03.09.2019

Gooskens C.; Heeringa W.: Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data, in: Language Variation and Change, Vol. 16/3, 2004

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Gopalakrishnan P. K.; Kar B.; Bose S. K.; Basu A.: Autoencoder-based Predictive Maintenance for IoT, in: IEEE Intelligent Systems, Vol. 19, 2019

Gudehus T.: Logistik 1, Springer, Berlin, Heidelberg, 2012

Gunay H. B.; Shen W.; Yang C.: Text-mining building maintenance work orders for component fault frequency, in: Building Research & Information, Vol. 47/5, 2019

Gupta V.; Lehal G. S.: A Survey of Text Mining Techniques and Applications, in: Journal of emerging technologies in web intelligence, Vol. 1/1, 2009

Heeringa W.: Measuring Dialect Pronunciation Differences using Levenshtein Distance, University of Groningen: Institute of Behavioral and Cognitive Neurosciences, 2004

Hevner A. R.; Salvatore T. M.; Ram S.; Park J.: Design Science in Information Systems Research, in: MIS Quarterly, Vol. 28/1, 2004

Holzinger A.; Stocker C.; Ofner B.; Prohaska G.; Brabenetz A.; Hoffmann-Wellenhof R.: Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data, Springer, Berlin, Heidelberg, 2013

Inmon W. H.: Data architecture, A primer for the data scientist: big data, data warehouse and data vault, Morgan Kaufmann, Waltham, 2015

Institut für Maschinelle Sprachverarbeitung: TIGER Corpus. URL: https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html, 09.09.2019

IoT Analytics: Predictive Maintenance Report 2019-2024, 2019

Janowski T.; Mohanty H.: Distributed computing and internet technology, 6th International Conference on Distributed Computing and Internet Technology, Bhubaneswar, India, 2010

Jiao J.; Zhang L.; Pokharel S.; He Z.: Identifying generic routings for product families based on text mining and tree matching, in: Decision Support Systems, Vol. 43/3, 2007

Jurafsky D.: Minimum edit distance. URL: https://web.stanford.edu/class/cs124/lec/med, 11.09.2019

Kao A.; Poteet S. R.: Text-Mining and Natural Language Processing, Introduction for the Special Issue, in: SIGKDD Explorations, Vol. 7/1, 2005

Kao A.; Poteet S. R.: Natural Language Processing and Text Mining, Springer-Verlag London Limited, London, 2007

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Karray H.; Chebbel-Morello B.; Zerhouni N.: A Formal Ontology for Industrial Maintenance, in: Terminology & Ontology, 2011

Khan F. I.; Haddara M. M.: Risk-based maintenance (RBM): a quantitative approach for maintenance/inspection scheduling and planning, in: Journal of Loss Prevention in the Process Industries, Vol. 16/6, 2003

Khoury R.; Karray F.; Kamel M.: Keyword extraction rules based on a part-of-speech hierarchy, in: International Journal of Advanced Media and Communication, Vol. 2/2, 2008

Klahold A.; Uhr P.; Ansari F.; Fathi M.: Using Word Association to Detect Multitopic Structures in Text Documents, in: IEEE Intelligent Systems, Vol. 29/5, 2014

Klein G.; Rowe S.; Decamps R.: JFlex lexical analyzer generator. URL: https://jflex.de/, 14.11.2019

Kobbacy K. A. H.; Murthy D. N. P.: Complex System Maintenance Handbook, Springer, London, 2008

Krizhevsky A.; Sutskever I.; Hinton G.: ImageNet Classification with Deep Convolutional Neural Networks, in: Advances in Neural Information Processing Systems, 2012

LanguageTooler GmbH: OpenThesaurus. URL: https://www.openthesaurus.de, 04.09.2019

Liu D.-R.; Ke C.-K.: Knowledge support for problem-solving in a production process: A hybrid of knowledge discovery and case-based reasoning, in: Expert Systems with Applications, Vol. 33/1, 2007

Lueth K. L.; Patsioura C.; Williams Z. D.; Kermani Z. Z.: Industrial Analytics 2016/2017, The Current State of Data Analytics Usage in Industrial Companies, in: IoT Analytics, 2016

Măndoiu I.; Zelikovsky A.: Bioinformatics research and applications, Proceedings of third international symposium of ISBRA, Atlanta, USA, 2007

Manning C. D.; Raghavan P.; Schütze H.: Introduction to information retrieval, Cambridge University Press, Cambridge, 2008

Martin M. R. W.; Reitz S.; Wehn C.: Kreditderivate und Kreditrisikomodelle, Eine mathematische Einführung, Vieweg & Sohn Verlag, Wiesbaden, 2006

Martinez A. R.: Natural language processing, in: Wiley Interdisciplinary Reviews: Computational Statistics, Vol. 2/3, 2010

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Matyas K.: Ganzheitliche Optimierung durch individuelle Instandhaltungsstrategien, in: Industriemanagement, 2002

Matyas K.: Instandhaltungslogistik, Qualität und Produktivität steigern, 7. Auflage, Hanser, München, 2019

Matyas K.; Nemeth T.; Kovacs K.; Glawar R.: A procedural approach for realizing prescriptive maintenance planning in manufacturing industries, in: CIRP Annals, Vol. 66/1, 2017

meliorate gmbh: The Optimal Maintenance Strategy. URL: http://meliorate.de, 20.05.2019

Mobley R. K.: An introduction to predictive maintenance, 2. Auflage, Butterworth-Heinemann, Amsterdam, New York, 2002

Moldaschl T.; Travnik A.; Frank A.; Bardong, Jochen: RFID-based Wireless Sensor Network for Semiconductor Manufacturing Equipment to enable Predictive Maintenance, 13th International Conference and Exhibition on Integration Issues, Barcelona, Spain, 2019

Mondal A.: Understanding Word vectors and Word2Vec. URL: http://www.stokastik.in, 16.09.2019

Morais A.; Barragués J. I.; Guisasola Aranzabal J.: Probability and statistics, A didactic introduction, CRC Press Taylor & Francis Group, Boca Raton, 2014

Mostafa S.; Dumrak J.; Soltan H.: Lean Maintenance Roadmap, in: Procedia Manufacturing, Vol. 2, 2015

Murff H.; Fitz-Henry F.; Matheny M.; Gentry N.: Automated Identification of Postoperative Complications Within an Electronic Medical Record Using Natural Language Processing, in: Journal of American Medical Association, Vol. 306/8, 2011

New Solutions GmbH: Finito Digital Shift Book. URL: https://www.schichtbuch.com, 18.11.2019

Newcombe R. G.: Confidence intervals for proportions and related measures of effect size, CRC Press, Boca Rato, 2013

Nguyen T.-S.; Nguyen L.-M.; Tojo S.; Satoh K.; Shimazu A.: Recurrent neural network-based models for recognizing requisite and effectuation parts in legal texts, in: Artificial Intelligence and Law, Vol. 26/2, 2018

Nicolas P. R.: Scala for machine learning, Leverage Scala and machine learning to construct and study systems that can learn from data, Packt Publishing, Birmingham, UK, 2014

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Nikfarjam A.; Emadzadeh E.; Muthaiyah S.; Mahadevan, V.: Text mining approaches for stock market prediction, The 2nd International Conference on Computer and Automation Engineering, Singapore, 2010

Nof S. Y.: Springer Handbook of Automation, Springer, Berlin, 2009

NSMG Ltd: AR smart googles for enterprise. URL: https://www.cbronline.com/in-depth/ar-goggles-for-enterprise, 22.10.2019

Pathak M. A.: Beginning data science with R, Springer, Cham, 2014

Pattanayak S.: Pro Deep Learning with TensorFlow, A Mathematical Approach to Advanced Artificial Intelligence in Python, Apress L. P, Berkeley, CA, 2017

Pawellek G.: Integrierte Instandhaltung und Ersatzteillogistik, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013

Phaedra Solutions. URL: https://www.phaedrasolutions.com/structured-data-vs-unstructured-data/, 26.07.2019

Plisson J.; Lavrac N.; Mladenic D.: A Rule based Approach to Word Lemmatization, University of Ljubljana: Department of Knowledge Technologies, 2004

Powers D.: Evaluation: From Precision Recall and f-Measure to ROC Informedness, Markedness & Correlation, in: Journal of Machine Learning Technologies, Vol. 2/1, 2011

Ram A.; Devaney M.; Qiu H.; Lee, Jay: Preventing failures by mining maintenance logs with case-based reasoning, 59th Meeting of the Society for Machinery Failure Prevention Technology, 2005

Rebala G.; Ravi A.; Churiwala S.: An Introduction to Machine Learning, Springer International Publishing, Cham, 2019

Reichel J.; Müller G.; Haeffs J.: Betriebliche Instandhaltung, Springer, Berlin, Heidelberg, 2018

Salo E.; McMillan D.: Value from free-text maintenance records, Converting wind from work orders into quantifiable, actionable information using text mining, University of Strathclyde: Department of Computer and Information science, 2018

Schenk M.: Instandhaltung technischer Systeme, Methoden und Werkzeuge zur Gewährleistung eines sicheren und wirtschaftlichen Anlagenbetriebs, Springer Verlag, Berlin, Heidelberg, 2010

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Sexton T.; Hodkiewicz M.; Brundage M.; Smoker, Thomas: Benchmarking for Keyword Extraction Methodology in Maintenance Work Orders, Annual conference of the prognostics and health management society 2018, Crawley, Australia, 2018

Shin J.-H.; Jun H.-B.: On condition based maintenance policy, in: Journal of Computational Design and Engineering, Vol. 2/2, 2015

Sipos R.; Fradkin D.; Moerchen F.; Wang Z., 2014: Log-based predictive maintenance, in: Proceedings of the 20th international conference on Knowledge discovery and data mining, editor: Macskassy S./Perlich C./Leskovec J./Wang W./Ghani R., ACM Press, New York, USA, pp. 1867–1876

Stanford NLP Group: Porter's algorithm. URL: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization, 14.11.2019

Stanford NLP Group: Stanford Tokenizer. URL: https://nlp.stanford.edu/software/tokenizer, 14.11.2019

Straka M.; Strakova J.: Tokenizing, POS Tagging, Lemmatizing and Parsing, Proceedings of the CoNLL Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, 2017

Studer R.; Benjamins V.R.; Fensel D.: Knowledge engineering: Principles and methods, in: Data & Knowledge Engineering, Vol. 25/1-2, 1998

Susto G. A.; Schirru A.; Pampuri S.; McLoone S.; Beghi A.: Machine Learning for Predictive Maintenance: A Multiple Classifier Approach, in: IEEE Transactions on Industrial Informatics, Vol. 11/3, 2015

Thiel K.; Kötter T.; Berthold M.; Silipo R.; Winters P.: Creating Usable Customer Intelligence from Social Media Data, Network Analytics meets Text Mining, University of Konstanz: Department of Computer and Information Science, 2012

Tschirk W.: Statistik: Klassisch oder Bayes, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014

Uhr P.: CIMAWA Entwicklung und Anwendung einer textbasierten Assoziations-Berechnungsmethode, Universität Siegen: Naturwissenschaftlich-Technische Fakultät, 2014

Ur-Rahman N.: Textual Data Mining For Knowledge Discovery and Data Classification: A Comparative Study, in: European Scientific Journal, ESJ, Vol. 13/21, 2017

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Ur-Rahman N.; Harding J. A.: Textual data mining for industrial knowledge management and text classification: A business oriented approach, in: Expert Systems with Applications, Vol. 39/5, 2012

Vezzetti E.; Marcolin F.: Similarity Measures for Face Recognition, Bentham Science Publishers, Sharjah, 2015

Wakiru J.; Pintelon L.; Muchiri P. N.; Chemweno P.: Maintenance Optimization: Application of Remanufacturing and Repair Strategies, in: Procedia CIRP, Vol. 69, 2018

Weiss S. M.; Indurkhya N.; Zhang T.: Fundamentals of Predictive Text Mining, Springer London, London, 2015

Williams T. P.; Gong J.: Predicting construction cost overruns using text mining, numerical data and ensemble classifiers, in: Automation in Construction, Vol. 43, 2014

Winkelmolen F.; Mascardi V.: Statistical Language Identification, University of Genova: Department of Computer Science, 2010

Yamada H.; Matsumoto Y.: Statistical dependency analysis with support vector machines, 8th International Conference on Parsing Technologies, 2003

Yoon C.; Dankel D.: Domain Specific Knowledge-based Information Retrieval Model using Knowledge Reduction, in: American Association for Artificial Intelligence, 2005

Yu L.; Wand S.; Lai K.: A Rough-Set-Refined Text Mining Approach for Crude Oil Market Tendency Forecasting, in: International Journal of Knowledge and Systems Sciences, Vol. 2/1, 2005

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

9 List of Figures

Figure 1: Global Market Development Predictive Maintenance 2
Figure 2: Relations between problem areas, objectives and research questions 6
Figure 3: Depiction of thesis structure 7
Figure 4: Basic measures of maintenance 9
Figure 5: Maintenance optimum 9
Figure 6: PF Interval of component lifecycle 13
Figure 7: Different data sources and categories 15
Figure 8: The basic model of knowledge-based maintenance 18
Figure 9: Knowledge-based Maintenance evolution 19
Figure 10: Examples of structured-, semi-structured and unstructured data 22
Figure 11: General framework of Text-mining 23
Figure 12: Example of tokenization 24
Figure 13: Examples of stemming 26
Figure 14: Model of artificial neuron 28
Figure 15: Heaviside step function 28
Figure 16: ANN with multiple layers 29
Figure 17: Recurrent neuron 30
Figure 18: Lemmatization by sequential recurrent neural networks 31
Figure 19: Word sequences with the highest joint probabilities 32
Figure 20: Visualization of all possible state transitions 33
Figure 21: Recurrent neural network for POS tagging 34
Figure 22: General architecture of Continuous Bag of Words 42
Figure 23: Cosine similarity measurement 43
Figure 24: Visualization of CIMAWA concept 44
Figure 25: Visualization of Associative Gravity Force 46
Figure 26: Model of a convolutional neural network for sentiment analysis 49
Figure 27: General model of a convolutional neural network 50
Figure 28: Distribution of negative and positive opinion words 52
Figure 29: Visualization of two different operation routings 57
Figure 30: Procedure of knowledge discovery process 58
Figure 31: Architecture of MKTPKS-based knowledge management and classification system 59
Figure 32: Model archiecture for knowledge extraction from maintenance logs 60
Figure 33: Overview of model architecture and methodology 61
Figure 34: Example of a failure protocol in original and target format 62
Figure 35: Procedural approach for prescriptive maintenance planning 63
Figure 36: Overall model architecture 64
Figure 37: Example for pre-processing of free text 65

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Figure 38: Procedural model for knowledge extraction 78
Figure 39: Steps of pre-processing, word embedding and keyword extraction 79
Figure 40: Tokenizing applied to maintenance reports 80
Figure 41: Example of abbreviation replacement 80
Figure 42: Example of punctuation removal 80
Figure 43: Language identification of maintenance reports 81
Figure 44: POS-tagging of maintenance reports 81
Figure 45: Lemmatizing of maintenance reports 82
Figure 46: Model of keyword extraction and tf-idf calculation 84
Figure 47: Co-occurence measuring of compounded terms 87
Figure 48: Schematic visualization of asymmetric AGF 88
Figure 49: Example of negative opinion word 89
Figure 50: Maintenance action asessment 90
Figure 51: Sentiment distribution of equipment L2301 90
Figure 52: RGB based coloring for equipment evaluation 91
Figure 53: AGF network of extracted key phrases 95
Figure 54: AGF and sentiment network of extracted key words and phrases 96
Figure 55: AGF and sentiment detection with various labels 97
Figure 56: Failure analysis based on AGF and sentiment detection 98
Figure 57: Linkage of human and object related knowledge assets 100
Figure 58: Text-mining methods structured according to purpose 101
Figure 59: Visualized instances of information 102
Figure 60: Possible example of a maintenance reporting form 105
Figure 61: Augmented Reality application in industrial maintenance 108

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

10 List of Formulas

Equation 1: Heaviside step function 27
Equation 2: Output value calculation of an artificial neuron 28
Equation 3: Markov probability equation 33
Equation 4: Formula of Bayes' theorem 35
Equation 5: Bayes' theorem based on word features 35
Equation 6: Likelihood under the assumption of independent features 35
Equation 7: Naive Bayes classification formula 35
Equation 8: Levenshtein distance matrix calculation 36
Equation 9: Last element of Levenshtein matrix 36
Equation 10: Logarithmic scaled term frequency 39
Equation 11: Binary frequency-based weighting 39
Equation 12: Local weighting adjustment based on document length 39
Equation 13: Global weighting adjustment 40
Equation 14: Term frequency - inverse document frequency weighting 40
Equation 15: Softmax function 42
Equation 16: Cosine similarity of two vectors 43
Equation 17: CIMAWA word association concept 43
Equation 18: Case differentiation of CIMAWA 44
Equation 19: Associative Gravity Force 45
Equation 20: Formula of Precision 71
Equation 21: Formula of Recall 71
Equation 22: F1-score formula 72
Equation 23: Matthews correlation coefficient 72
Equation 24: Cohen's kappa coefficient 72
Equation 25: Formula of observed accuracy 72
Equation 26: Formula of expected accuracy 72
Equation 27: Estimator of a proportion 73
Equation 28: Confidence interval based on significance level 73
Equation 29: Width of confidence interval 74
Equation 30: Estimated sample size based on the confidence interval 74
Equation 31: Implemented term frequency concept 83
Equation 32: Implemented inverse document frequency concept 83
Equation 33: Implemented Levenshtein ratio inequality 87
Equation 34: Concept of CIMAWA and AGF 87
Equation 35: Sample size calculation for model evaluation 91
Equation 36: Observed accuracy of applied sentiment analysis 92
Equation 37: MCC value of sentiment analysis evaluation 92
Equation 38: Expected accuracy of applied sentiment analysis 92
Equation 39: Cohen's Kappa of sentiment analysis evaluation 92

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

11 List of Tables

Table 1: Rules and examples of Porter Stemmer 25
Table 2: Measure based rule for stemming 25
Table 3: Possible POS-tags of word examples.....32
Table 4: Example of Levenshtein matrix.....37
Table 5: Un- and pre-processed example sentences for BoW representation 38
Table 6: Example of Term-Document matrix 38
Table 7: Example of tf-idf weighting.....40
Table 8: Misclassification and sensitivity rates of customer churn prediction 53
Table 9: State-of-the-art content overview.....66
Table 10: Evaluation methods and categories.....69
Table 11: Confusion matrix.....71
Table 12: Example of maintenance records 75
Table 13: Extract of maintenance records represented by their keywords 85
Table 14: Classification branches of opinion words.....90
Table 15: Confusion matrix of regarded evaluation sample.....92
Table 16: Summary of detected rules and association patterns 99
Table 17: Summary of proposed research directions for model enhancement.....108

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

12 List of Abbreviations

Acronym	Description
3D	Three-dimensional
ADJD	adjective
AGF	Associative Gravity Force
AI	Artificial Intelligence
ANN	Artificial Neural Network
APC	Automatic Process Control
APPR	left hand preposition
AR	Augmented Reality
BOW	Bag of Words
CBOW	Continuous Bag of Words
CBR	Case Based Reasoning
CIMAWA	Concept for the Imitation of the Mental Ability of Word Association
CRISP-DM	Cross Industry Standard Process for Data Mining
DFA	Deterministic Finite Automation
e.g.	exempli gratia
EDP	Electronic Data Processing
et al.	et alii
etc.	et cetera
FN	False Negative
FP	False Positive
GUI	Graphical User Interface
IoT	Internet of Things
IT	Information Technology
JSON	Java Script Object Notation
KBM	Knowledge-Based Maintenance
KPI	Key Performance Indicator
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NN	noun
Nr.	Number
o.E.	ohne Ergebnis
OEE	Overall Equipment Effectiveness
OWL	Web Ontology Language
PdM	Predictive Maintenance
PF	Potential-to-functional Failure
POS	Part-of-Speech
RCM	Reliability-Centered Maintenance
RDF	Resource Descriptive Framework
ReLu	Rectified Linear unit
RFID	Radio-Frequency Identification
RGB	Red-Green-Blue
ROI	Return of Investment
RUL	Remaining Useful Life

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Acronym	Description
sim.	Simulation
SQL	Structured Query Language
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TF-IDF	Term frequency - inverse document frequency
TLC	Total Life-cycle Cost
TN	True Negative
TP	True Positive
TPM	Total Productive Maintenance
VR	Virtual Reality
VVPP	past participle of verb
XML	Extensible Markup Language

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar.
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub