Applied Artificial Intelligence

An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Business Analytics in Strategic Purchasing: Identifying and Evaluating Similarities in Supplier Documents

Frank Bodendorf, Benedict Wytopil & Jörg Franke

To cite this article: Frank Bodendorf, Benedict Wytopil & Jörg Franke (2021) Business Analytics in Strategic Purchasing: Identifying and Evaluating Similarities in Supplier Documents, Applied Artificial Intelligence, 35:12, 857-875, DOI: 10.1080/08839514.2021.1936423

To link to this article: https://doi.org/10.1080/08839514.2021.1936423

Published online: 19 Jul 2021.

Submit your article to this journal

Article views: 912

View related articles

View Crossmark data

Citing articles: 1 View citing articles

APPLIED ARTIFICIAL INTELLIGENCE
2021, VOL. 35, NO. 12, 857–875
https://doi.org/10.1080/08839514.2021.1936423

Business Analytics in Strategic Purchasing: Identifying and Evaluating Similarities in Supplier Documents

Frank Bodendorf[a], Benedict Wytopil[b], and Jörg Franke[a]

[a]Institute for Factory Automation and Production Systems, Friedrich-Alexander-University of Erlangen-Nuremberg (FAU), Erlangen, Germany; [b]IPRI International Performance Research Institute, Stuttgart, Germany

ABSTRACT

The increasing digitalization in the automotive industry is influencing the structure of the traditional value chain and calls for the handling of large amounts of data to remain competitive in a constantly changing environment. This results in new challenges for purchasing management, which has to cope with agile integration of service providers as well as interorganizational process automation using electronic data exchange platforms. This work analyzes the electronic document stream on supplier management platforms by proposing an automated text mining framework. Both textual components, e. g., requests for information and offers, and narrative material, e. g., financial and calculation data, are being analyzed by topic modeling and descriptive statistics. The methodological approach is introduced and illustrated by the use case of service provider documents in purchasing processes. The results reveal financial potential for purchasing and generally contribute to supply chain cost management.

ARTICLE HISTORY
Received 05 October 2020
Accepted 19 October 2020

Introduction

In order to respond appropriately to current challenges in the automotive industry, such as the increasing complexity and variability of product structures, shorter product life cycles (Delhi 2016), the growing competitive pressure from start-ups and technology companies (Paradkar, Knight, and Hansen 2015), and the increase in the share of external value added, not only flexible production systems but also highly efficient processes in other business areas like purchasing and supply chain management are required (Hung 2006) (Porter 1985). purchasing and especially cost management as a sub-function plays a crucial role in business practice. It has become a critical success factor in terms of ensuring sustainable competitiveness (Orina 2018). As a result of the restructuring of the automotive value chain the average vertical integration of Original Equipment Manufacturers (OEMs) has been reduced from 35% in

CONTACT Frank Bodendorf frank.bodendorf@faps.fau.de Institute for Factory Automation and Production Systems, Friedrich-Alexander-University of Erlangen-Nuremberg (FAU), Egerlandstraße 7-9, 91058, Erlangen, Germany

1990 to approximately 20% in 2015 (Olausson, Magnusson, and Lakemond 2009) (Statista 2010). Development service providers (DSPs) are becoming increasingly important here.

DSPs are an important competence leader in the global (7%) and especially in the German (12%) automotive R&D value chain (Kleinhans and Bräuning 2015). The majority of OEMs, but also some suppliers rely on the development capabilities of DSPs. In 2012 DSPs generated 54.8% of sales with OEMs and a further 8.2% with automotive suppliers (Kleinhans and Bräuning 2015). DSP services range from product development (individual components and subsystems as well as entire modules and overall systems) to process development (manufacturing processes or the design of tools and entire systems) and product and process development support activities (project management, documentation, costing and quality assurance measures) (see Figure 1) (Reichuber 2010). OEMs and suppliers benefit in many ways from the integration of DSPs into the automotive R&D value chain. In addition, DSPs offer a high degree of flexibility and can therefore provide timely support during peak demand periods in development projects. In addition, DSPs offer a high degree of flexibility and can therefore provide timely support during peak demand periods in development projects.

Another advantage is the cost structure of most DSPs. These usually have lower overheads as a result of the smaller company size and the more flexible working time models. Increasingly, DSPs also contribute to reducing

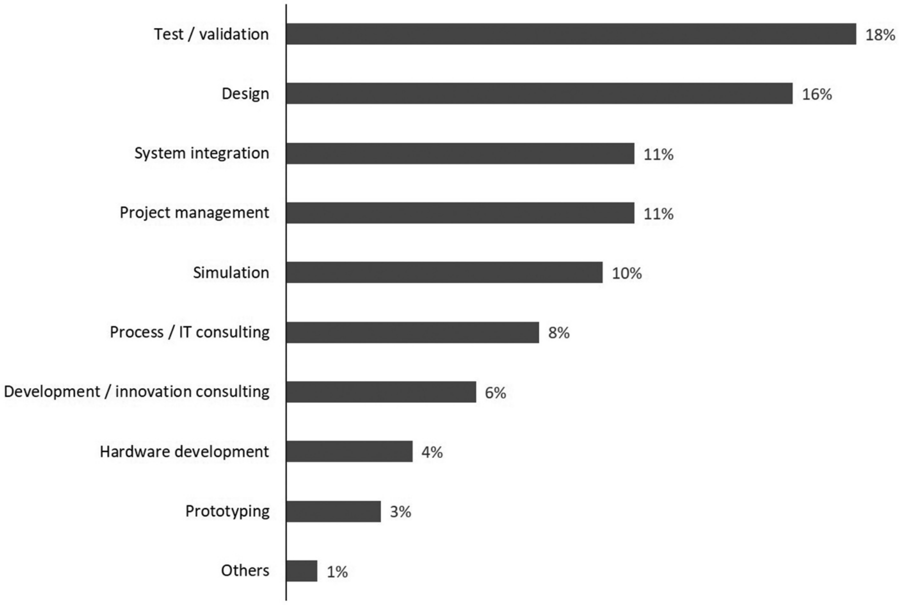

Figure 1. DSP Portfolio.

development risks, which consequently leads to a reduction of corresponding risks for the customer (Blöcker 2016)(Kleinhans and Bräuning 2015).

The high level of DSP integration together with a large amount of information exchanged (e. g., contracts) confronts the OEM with challenges of complexity reduction as well as information and cost management. The first reaction to these challenges is electronic data interchange (EDI), which enables a relatively high degree of integration and transparency between service providers and OEMs. In theory, Cooper (2017) and Cooper and Slagmung (2004) shows how companies can operate interorganizational data management by EDI in order to gain cost transparency and clarity in the complex transactions of supplier-manufacturer relationships. Kauffman and Mohtadi (2004) analyze the organizational adoption behavior of EDI in purchasing (Kaufmann and Mohtadi, 2002). Motivated by the high importance of DSP in the automotive industry combined with a high relevance in improving efficiency of business processes as well as the growing acceptance of IT solutions (e. g., EDI), tools of data science – in particular methods of data mining (DM) and machine learning algorithms (ML) are of growing importance (Accorsi and Stocker 2012) (Bose and Mahapatra 2001)(Grigori et al. 2004) (Gupta 2019). Advances in algorithms and the available computing power as well as an exponentially increasing amount of data have significantly driven the performance of data analytics (Henke et al.,). Here, Natural Language Processing (NLP), which enables the extraction of information from textual data sets is a crucial sub-area (Cambria and White 2014) (Salama and El-Gohary, 2016).

An evaluation of the n-tier EDI system in a given purchasing context of a car manufacturer suggests that in the past, multiple use of the same or similar development services from third parties have occurred, resulting in unnecessary expenses.

The underlying research project of this paper aims to investigate similarity structures in DSP contract documents. A NLP machine learning model has been developed that uses historical data to analyze the similarity between documents like calculations, offers or tenders.

Table 1 summarizes fields of application and corresponding objectives. Even if the expectations are mostly formulated qualitatively due to the mostly exploratory character of the applications, the analyses should have predominantly quantitative components.

Table 1. Text mining and corresponding research objectives.

Data dimension	Field of action	Objective
textual, descriptive	topic definition	identification of topics
textual, descriptive	duplicate removal	identification of duplicates
textual, descriptive	similarity assessment of documents in purchasing	identification of documents that could contain similar development services

In Section 2 the background of topic modeling in business analytics is introduced. After giving this theoretical background Section 3 focuses on outlining methodologies for each field of application to illustrate corresponding implementations with real data in Section 4. After a report of key findings Section 5 discusses implications for practitioners and managers and shows limitations as well as potentials for further development.

Text Mining Potentials

While applications of text mining for business analytics are abundant in the literature, there is scant evidence demonstrating the use of topic modeling. In this section we point toward different research streams concerning language-based business analytics. These vary according to (a) the text source and (b) the analytical model, i. e., supervised/unsupervised learning. Text mining often yields new insights from narrative language for steering the organizational decision-making and operation of firms. One characteristic feature is the source of the written materials. They often consist of user-generated content, such as product reviews or social media postings, where sentiment analysis facilitates insights into the opinion of customers toward products on social media like twitter or facebook (Mostafa 2013). The gender affiliation for example can be recognized based on blinded e-mail text documents (Berezina et al. 2016).

Alternatively, investors can investigate whether narrative disclosures in annual financial reports of firms contain value-relevant information for predicting market performance (Balakrishnan, Qiu, and Srinivasan 2010) (Paradkar, 2010) (Kearney and Liu 2013). Or committees can evaluate the submissions to crowdsourcing websites in order to select winners and adjust monetary reward (Walter and Bac, 2013).

Moreover, internal use of text and language reflects the structure and processes of organizations. So, analytics can be applied in purchasing not only to prevent and detect fraud. Moreover, data mining techniques such as text mining and cluster analysis can be used to improve visibility of purchasing patterns and provide decisionmakers with insight to develop more efficient sourcing strategies, in terms of cost and effort. Furthermore, they can help in storing and managing purchasing contracts (Tan and Lee 2015).

In addition to that a study shows a procedure for automated learning of negotiation strategies and stylized business negotiations (Oliver 1996).

Further application potentials can be found in business process management, for example to check validation and plausibility of existing processes or to transform continuous text into process models (Leopold, Mendling, and Polyvyanyy 2014)(Leopold, Pittke, and Mendling 2014). For example, violations of process models or compliance guidelines can be detected and documented (Accorsi and Stocker 2012)(Van der Aalst, 2010).

The applications of text mining also show considerable variations in terms of the underlying methods. On the one hand, a lot of use cases require supervised learning with a priori labels. Examples include automated assignment of IT tickets to the appropriate service unit (Goby et al. 2016), forecasts of news-based stock price changes (Pröllochs, Feuerriegel, and Neumann 2016), and predicting users' affect (Rao et al. 2016). Others rely on unsupervised methods, such as clustering or topic modeling, which are able to shed light on the patterns within business data. Illustrative demonstrations include measuring business proximity (Shi, Lee, and Whinston 2016), predicting interest among tourists (Brandt, Bendler, and Neumann 2017), and forming IT support groups on the basis of the content of helpdesk tickets (Goby et al. 2016).

Research Framework

Theoretical Background

The amount of unstructured data clearly outweighs the amount of structured data. Experts put the share of unstructured data at 70–90% (Subramaniyaswamy et al. 2015). Apart from images and sound recordings, this mainly involves textual data (Baars and Kemper 2008). Accordingly, many organizations and companies have large amounts of textual data. In order to generate added value from this data, two major challenges must be met. First, an effective system for data storage and data management is necessary. Second, there must be efficient algorithms to process and analyze textual data to extract useful information (Holzinger and Pasi 2013). Most ML algorithms are designed for numerical data. Therefore, special methods and techniques for algorithmic processing and transformation of natural language have formed under the term “Natural Language Processing” (NLP). While regular structures can often be found in numerical data sets, textual data sets usually do not follow a regular syntax, are therefore very variable and cannot be directly analyzed by classical statistical models. In the field of NLP and text analytics often basic forms of ML are used, but specific algorithms sometimes differ considerably (Manning, Raghavan, and Schuetze 2018)

(Sarkar 2016) (Sarkar, Bali, and Sharma 2018). In the following, three typical instruments relevant for this work are discussed.

Topic Modeling is a method for analyzing the distribution of semantic word groups in text collections, so-called "Topics." It is suitable both for the explorative examination of the contents of a corpus and for obtaining features for computer-aided text classification. The procedure does not require external dictionaries, training data or similar and works in principle independently of language or orthographic conventions. Only the frequencies of characters at word level are statistically examined and translated into presumed semantic relationships. This makes Topic Modeling a particularly flexible method with regard to its requirements for text type and text quality.

Latent Dirichlet Allocation (LDA) is a probability model for a corpus (e. g., a collection of text documents). The basic idea is that each document consists of a number of topics are not directly visible, i. e. latent. Furthermore, each topic is available as a mixture of words forming the theme. In this model, the different words (and ultimately documents) of a corpus with the highest possible probability of a topic can be detected. The assigned topics form the later clusters.

Based on the assignment of words and documents to topics (and thus clusters), the topic composition of a document can be determined (for example: 20% topic A, 70% topic B and 10% topic C). In addition, you can use the determining keywords per cluster (e.g. the most frequently used words) which approximate the content of a topic (or cluster). Basically, the topic assignment of LDA is based on a learning procedure which is based on Bayesian statistics and can be assigned to the methods of unsupervised learning. Furthermore, the basic idea is a Bag-Of-Words approach, which only allows a document to be collection of words, but without semantics.

Similarity Analysis between text bodies is done by determining the lexical or semantic "proximity" between corresponding texts. In Table 2, the difference between lexical and semantic similarity is illustrated by two examples. While lexical similarity is based on the words used, semantic similarity takes into account the deeper meaning and context

Table 2. Illustration of lexical and semantic similarity.

	lexical	semantic
The cat chases the mouse. The mouse chases the cat.	very similar	not similar
The queen has died. The king's wife is dead.	not similar	very similar

of the text section. Great progress has been made in methods for evaluating semantic similarities thanks to models such as the "Universal Sentence Encoder" or ELMo (Embeddings from Language Models), which have been trained with up to one billion words. Nevertheless, the use of these methods involves a great deal of effort when dealing with multilingual or non-English texts. Furthermore, semantic similarity is particularly important when analyzing coherent phrases or entire sentences. (Cer et al. 2018) (Peters, 2013). Consequently, the research focuses on lexical similarity.

The lexical similarity typically does not take into account the actual meaning of the words or the whole phrase in context. However, this does not mean that the use of this form of similarity cannot be effective. Algorithms for evaluating lexical similarity are used for clustering documents as well as for removing redundant text or test components and duplicates in databases.

Developmental Framework

With CRISP-DM (Cross Industry Standard Process for Data Mining) and the Standard ML Pipeline, the procedure of data mining and predictive modeling is set on a methodological basis. Besides company-specific process models, the KDD (Knowledge Discovery in Databases) and SEMMA (Sample, Explore, Modify, Model, and Access), the CRISP-DM approach is most widely used (Azevedo and Santos 2008) (Wirth and Hipp 2000). The CRISP-DM and the Standard ML Pipeline can be regarded as complementary procedure models (Piatetsky 2019). CRISP-DM has a broader focus, both on the actual process and on the capabilities of the tools used (ML algorithms, statistical and data mining software, business intelligence concepts). The standard ML Pipeline, on the other hand, focuses on the procedure for creating ML models. The understanding of data and business is considered a prerequisite here. The methodology used for developing the ML models outlined in this paper is based on these two complementary procedural models.

Method Overview

The similarity assessment of documents is carried out according to the basic procedure shown in Figure 2. The individual steps of the process are based on the frameworks mentioned in Section 3.2 and follows the NLP methods motivated by Manning and Pröllochs (Manning,

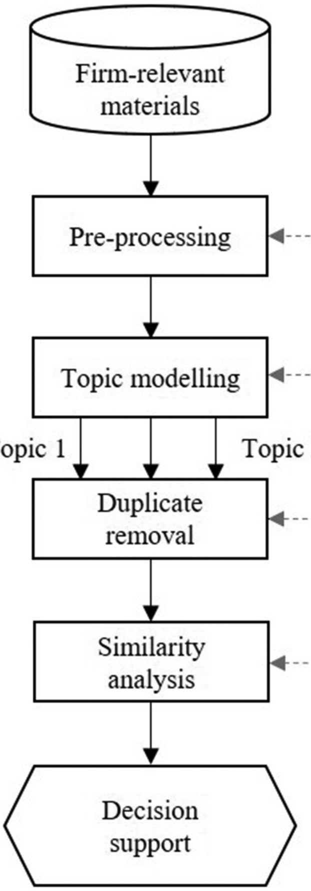

Figure 2. Process of similarity assessment.

Raghavan, and Schuetze 2018)(Manning and Schütze 1999) (Pröllochs and Feuerriegel 2020).

The preprocessing encompasses common operations in text mining and is supplemented by the removal of duplicate documents (Pang and Lee 2008)(Ravi K. and Ravi V., 2015) and combinable with optional filtering steps (e. g. years, business area). Duplicates are to be understood in the sense of identical or easy adaption of the same document (e. g., new version of a document). For preprocessing, the following preparatory steps must be carried out first:

•The related texts must be broken down into individual words (tokenization). The resulting list can be processed quickly and easily and forms the basis for subsequent modeling.

•A list must be created that contains words from all documents. This process is necessary to assess the word significance in a document, considering the context of all words.

The preparation includes transformation operations and the removal of character strings. In concrete terms, the following steps must be performed (Sarkar 2016) (Sarkar, Bali, and Sharma 2018):

•Implementation of consistent lower case (i. e., all characters are lower case) to reduce the amount of data to be processed. Identical words that differ in their capital case (e. g., project/PROJECT) are otherwise regarded as two different words.

•Removal of numbers and special characters, as they do not provide any added value in the lexical similarity analysis.

•Removal of words that have no informative value (so-called stop words) (Gloeggler 2003).

•Stemming to reduce different morphological forms of a word to their common stem. The Porter-Stemmer algorithm, for example, can be used here (Sparck Jones and Willett 1997).

LDA modeling aims at creating a similarity matrix, which can then be used to identify duplicates. Various algorithms are suitable for implementing this. The LDA (Latent Dirichlet Allocation) algorithm is widely used for comparable problems (Chen and Zhang 2014) (Chen, 2017) (Hutchison, Kanade, and Kittler 2013) (Rus, Niraula, and Banjade 2013) (Towne, Rosé, and Herbsleb 2016)(Zhu and Li 2012). Especially for the comparison of texts that are not too short (> 40 tokens) this algorithm is convincing (Hong and Davison 2010). The LDA algorithm is an unsupervised generative algorithm that assigns specific topics to the documents to be evaluated, which in turn are composed of significant words. Subsequently, topics with a certain weighting are assigned to each document (Blei 2012). Based on the specific combination of topics and weighting, a similarity matrix can now be generated (see Table 3).

Table 3. Similarity matrix.

	Document 1	Document 2	Document 3
Document 1	1	0.5	0.75
Document 2	0.5	1	0.25
Document 3	0.75	0.25	1

1) Duplicate Removal

The similarity matrix, based on the topics generated by the LDA model, shows the pairwise similarities, using the similarity score from the gensim library (Vorontsov, 2015). The similarity score is based on the cosine similarity, which is a wildly used method to compare two term-frequency vectors. Formula 1 illustrates the general concept of the cosine similarity where x and y are two "topics-weighting-vectors" for comparison.

$$\text{sim}(x,y) = \frac{x*y}{\sqrt{x_1^2 + x_2^2 + \ldots} * \sqrt{y_1^2 + y_2^2 + \ldots}} \quad (1)$$

We use the "topics-weighting-vectors" generated by the LDA model as an input for the pairwise cosine similarity calculation.

The similarity score is a number between 0 (no similarity) and 1 (equal) and gives a quantitative evaluation of the similarity based on the topics and weights defined by the LDA algorithm and assigned to the respective documents. As very similar documents are regarded as duplicates, the definition of a similarity threshold is very important. However, it is not possible to determine an optimal similarity score. As a consequence, different similarity ranges are simulated. The lower limits are usually set to values between 0.8 and 0.99 (similar to very similar). The upper limit is always 1 (identical). A qualitative random simulation may lead to a lower limit value of 0.975, i. e., all documents with a similarity score of more than 0.975 to another document are considered duplicates (which can be validated by random sampling) and removed.

2) Similarity Analysis

The similarity matrix based on the topics generated by the LDA model shows the pairwise similarities by means of the similarity scores. Documents that show a high degree of similarity are then selected. First, it is examined qualitatively whether a similarity really exists or whether it is an updated version of the same document. For this purpose, a list of similar documents is generated for various lower limits. This pre-selection of different document clusters helps a lot to speed up the detection and removal of "real" duplicates and "very similar" documents, which is done manually.

Case Study

Study Setup

The study takes place at the purchasing department of an Original Equipment Manufacturers (OEM). OEMs represent the focal point of the value chain and combine their own, a combination of their own and externally sourced or purely externally sourced resources. These may include production services (raw materials, individual parts, components and modules), software development and other services. The final product is sold to the end customer on the market under their own brand name.

The above-mentioned task of sourcing is done by the purchasing department.

A typical purchasing process begins with the need to define buying requirements based on the demands of the firm's final customer. Once the specifications have been developed, a buying team led by the purchasing and supply manager will prequalify suppliers, generate requests for proposals, evaluate the proposals, and select a supplier based on established selection criteria. Contract negotiations result in the terms and conditions of a formal contract. Ordering routines and transaction-processing guidelines are established for all purchases that take place under the umbrella of the negotiated contract. Closing the loop is a supplier evaluation system that assesses supplier performance that provides information to be used as the basis for rating the supplier (e.g., excellent, good, fair, unacceptable).

This research focuses on the contract negotiations phase. For text analytics we chose EDI supplier platforms and have a closer look on historic semi structured data in the form of quotes and cost calculation reports that are not accessible for traditional numeric search algorithms. In addition, they contain a lot of information which are hard to access.

Dataset

The database consists of a folder structure that serves as a document repository for purchasing over the period from 2005 to 2019. It is not a complete data set (i. e. not covering all purchasing projects). The folder structure contains 85,513 files which are made up of:

- Request for Information (RFI) data aggregating information from different suppliers prior to formally sourcing products or services,
- Request for proposal (RFP) data consisting of detailed and comparable proposals from different suppliers for a defined product or service,
- Request for Quotation (RFQ) data bundle documents used when inviting suppliers and subcontractors to submit a bid on projects or products. An RFQ is suitable for sourcing products that are standardized or produced in repetitive quantities. A technical specification must be provided as well as commercial requirements,
- financial and cost calculation data (e. g. financial statements or cost breakdown structurers) in various formats (MS Excel, MS Powerpoint,

MS Word, MS Outlook, PDF) that were created and exchanged with DSPs in the context of purchasing projects.

The following example focuses on Excel files. To give a basic impression of the database, Table 4 shows the number of files and lines per year. There are 49,758 files with a total of 90,481,910 lines available for further processing.

Table 4. Quantitative distribution of the textual datasets.

Year	Number of Lines	Share [%]
2005	62,832	0%
2006	89,584	0%
2007	254,887	0%
2008	1,124,570	1%
2009	475,443	1%
2010	560,434	1%
2011	14,633,428	16%
2012	2,526,757	3%
2013	3,212,047	4%
2014	5,967,968	7%
2015	4,068,857	4%
2016	10,702,953	12%
2017	9,511,064	11%
2018	37,268,892	41%
2019	22,194	0%
Total	90.481.910	100

Findings

1) Identification of Topics

To run the LDA, it is necessary to define a certain number of subjects to be identified. In this use case the subjects are DSP related topics in a purchasing context.

Not all topics can be separated clearly. In the experimental setting a number of 100 topics has proven to be optimal. A number suggested in the literature is approx. 20 (Blei, 2012; Niederhoffer 1971; Ramage, 2009). Given the data volume of close to 100 million lines a set of 100 topics does not seem to be large. The topic names are defined by domain experts based on the individual terms in the topics list (see Table 5).

For example, the terms “derivatives,” “diesel,” “hybrid,” “charging system” and “electric drive” are summarized as the topic named “drive technology.”

To increase the informative value of the topics, the Term Frequency – Inverse Document Frequency approach (TF-IDF) is used to assign a weight based on their frequency and thus significance for the respective topic.

Table 5. Excerpt of extracted topics and weights.

Vehicle module development.	development	0.596	concept design	0.066	team management	0.046	rrnm	0.016	story	0.014	simulation	0.014	control system	0.012	belt	0.010
Drive technology.	*derivative*	0.397	*diesel*	0.069	*hybrid*	0.056	*confirmation of function*	0.044	*concept confirmation*	0.029	*charging system*	0.027	*petrol*	0.020	*electric drive*	0.017
Vehicle modules.	*complete*	0.364	*coverage*	0.108	*instrument panel*	0.071	*front panel*	0.031	*rear window*	0.025	*bolt*	0.017	*windlauf*	0.017	*decor*	0.016
Modeling.	*specifically*	0.335	*swip*	0.074	*micronova*	0.033	*labview*	0.028	*application*				*development*	0.015	*mechanical variant*	0.011
software integration	0.010	*sensor*							*technology design*	0.007						
Engineering service provider.	*gmbh*	0.262	*engineering*	0.068	*engineering office*	0.032	*technology partner*	0.026	*engineering partner*	0.024	*system*	0.023	*on-call order*	0.019	*group*	0.017
Construction costs.	*series project*	0.224	*catia*	0.180	*mixed hourly rate*	0.132	*project*		*coordination*	0.105			*organization*	0.056	*scope of components*	0.026
analysis	0.022	*takeover*	0.018													
Bavarian locations.	*location 1*	0.210	*location 2*	0.193	*location 3*	0.142	*location 4*	0.044	*material*	0.037	*unpack*	0.035	*pack*	0.034	*drawing feature*	0.018
Development service provider.	*on-call order*	0.193	*bavaria*	0.041	*system supplier*	0.034	*germany*	0.033	*location 4*	0.032	*vehicle*		*technology*	0.031	*design*	0.029
operator model	0.029															
External orders.	*bertrandt*	0.189	*on-call order*	0.147	*engineering partner*	0.084	*ergonomics*	0.066	*procurement*	0.056	*engineering partner*	0.045	*fraunhofer*	0.036	*type testing*	0.035
Software tests.	*test case*								*preparation*	0.148	*microfuzzy*	0.049	*test case execution*	0.047	*software release*	0.040
test platform support	0.040	*error*	*solution*	0.037	*vehicle prototype*	0.027	*test case validation*	0.027								
Concept validation.	*architecture*	0.129	*project*	0.106	*scope*	0.085	*concept*	0.067	*system supplier*	0.039	*geom*	0.031	*requirements management*	0.029	*electrification*	0.020
Simulation.	*simulation*	0.121	*variant*	0.084	*modelling*	0.057	*analysis*	0.053	*try*	0.029	*crash*	0.029	*calculation*	0.028	*component*	0.026
Climate tests.	*klimakamm*	0.087	*application*	0.076	*climate chamber*	0.038	*test engineer*	0.030	*testing technology*	0.028	*print*	0.028	*actuators*	0.025	*closed*	0.024
System test.	*test*	0.081	*system*	0.029	*conducted*	0.023	*overall*	0.017	*measurement*	0.016	*safety*	0.015	*report*	0.015	*function*	0.013
Drive technology.	*drive*	0.058	*engine*	0.051	*faar*	0.037	*electrical*	0.021	*flexray*	0.019	*transmission*	0.018	*sensor*	0.017	*internal combustion engine*	0.017

The naming of the topics is not mandatory for the next step, but it can represent an added value for itself.

2) Removal of Duplicates

After having defined the topic list, duplicate documents can be identified with the help of similarity scores. In the use case the amount of data for the subsequent similarity assessment is reduced from 49,758 to 5,122 documents.

3) Similarity Assessment of Documents in Purchasing

Finally, the pairs of similar documents are evaluated. This is first carried out qualitatively to determine whether a similarity really exists or whether it is an updated version of the same document. For this purpose, a list of similar documents is generated for various lower limits. Here the lower limits for the similarity scores is set to values between 0.9 and 0.999 (see Table 6).

Due to the large number of documents for the individual classes that are identified as similar, individual pairs are examined randomly to provide a basis for discussion of the procedure. This research follows the procedure shown in Figure 3. As a result, all documents that are similar are identified.

Table 6. Number of documents classified as "similar" (total), average number of "similar" classified documents.

	Total	Avrg.	Share
SimScore > 0.9	233.552	45.6	0.7
SimScore > 0.925	180.132	35.2	0.62
SimScore > 0.95	120.684	23.6	0.53
SimScore > 0.975	57.182	11.2	0.39
SimScore > 0.99	18.696	3.7	0.25
SimScore > 0.999	1.036	0.2	0.08

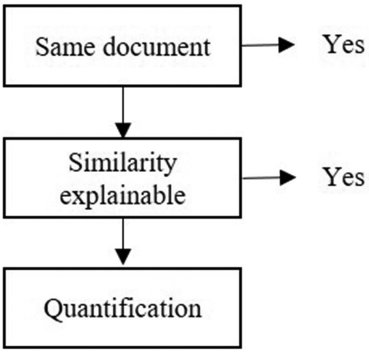

Figure 3. Similarity assessment procedure.

Conclusions

Results

In the course of the methodology described, it was first necessary to remove duplicates in a large semi-structured data set. The removal of duplicates by means of a similarity matrix generated by an LDA model has proven to be effective on a methodological level as well as in practical implementation. When applying this methodology, the size of the data set was reduced from 49,758 to 5,122 documents. The effectiveness of this duplicate removal is validated by random sampling. The similarity score can be adjusted in such a way that it is optimized with regard to sensitivity (low similarity score) or specificity (high similarity score). The similarity score was successfully applied to a real data set and similar documents were found in really "big data."

Furthermore, the effectiveness of the algorithms based on theme modeling (here: LDA) could be confirmed. In this field of application, the potential of optimizing the hyper parameters could also be shown. The coherence of the model, as a performance indicator for this form of NLP algorithms, could be improved by parameter optimization from −14.27 to −1.91 (optimum: coherence = 0).

Business Implications

This work particularly demonstrates how advanced analytics can provide business value. Bennett and Lemoine (2014) analytics help firms to minimize their costs in order to stay competitive in a VUCA-environment. Some even claim for a merge of IT solutions and business strategies with big data being one element (Bharadwaj, 2013).

Our work contributes to information systems research by showing to purchasing management and controlling departments that analytic tools are able to bring together people, tasks, and technology to optimize the company value. Additionally, the considered use case shows the gain of additional information extracted from own company data and not pure user data (Martens et al. 2016) (Qi et al. 2016). It is remarkable that current research on the analysis of company owned data seems to be underrepresented in big data analytics (Chen, 2017).

Considering the research results, managers should push predictive applications that work with textual data. In this way, historical knowledge can be used to allocate documents and cost calculations in their creation process to historical document clusters. This assignment can support document creation by making the new document plausible and thus consistent with historical documents. In addition, such identification of similar historical documents can help with the quantitative evaluation of certain cost items in the context of cost engineering.

References

Accorsi, R., and T. Stocker, "On the exploitation of process mining for security audits, The 27th annual ACM Symposium on Applied Computing, New York, 2012.

Azevedo, A., and M. F. Santos. 2008. "*KDD, SEMMA and CRISP-DM: A parallel overview.*" IADIS European Conference on Data Mining, Amsterdam, Netherlands.

Baars, H., and H. G. Kemper. 2008. "Management support with structured and unstructured data—an integrated business intelligence framework. *Information Systems Management* 25 (2):132–148. doi:10.1080/10580530801941058.

Balakrishnan, R., X. Y. Qiu, and P. Srinivasan. 2010. "On the predictive ability of narrative disclosures in annual reports. *European Journal of Operational Research 202(3)* 202 (3):789–801. doi:10.1016/j.ejor.2009.06.023.

Bennett, N., and G. J. Lemoine. 2014. "What a difference a word makes: Understanding threats to performance in a VUCA world. *Business Horizons* 57 (3):311–317. doi:10.1016/j.bushor.2014.01.001.

Berezina, K., A. Bilgihan, C. Cobanoglu, and F. Okumus. 2016. "Understanding satisfied and dissatisfied hotel customers: Text mining of online hotel reviews. *Journal of Hospitality Marketing & Management* 25 (1):1–24. doi:10.1080/19368623.2015.983631.

Bharadwaj, A., O. A. El Sawy, P. A. Pavlou, and N. Venkatraman. 2013. "Digital Business Strategy: Toward A Next Generation of Insights. *MISQ* 37 (2):471–482.doi:10.25300/MISQ/2013/37:2.3

Blei, D. M. 2012. "Probabilistic topic models. *Communications of the ACM* 55 (4):77. doi:10.1145/2133806.2133826.

Blei, D. M., A. J. NG, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Blöcker, A. 2016. Branchenanalyse entwicklungsdienstleister. In *Working paper forschungsförderung No. 017*. Düsseldorf: Hans-Böckler-Stiftung, p.43. https://www.econstor.eu/bitstream/10419/215949/1/hbs-fofoe-wp-017-2016.pdf

Bose, I., and R. K. Mahapatra. 2001. "Business data mining—a machine learning perspective. *Information & Management* Elsevier, Amsterdam. 39 (3):211–225. doi:10.1016/S0378-7206(01)00091-X.

Brandt, T., J. Bendler, and D. Neumann. 2017. Social media analytics and value creation in urban smart tourism ecosystems. *Inf Manag* 54 (6):703–713. doi:10.1016/j.im.2017.01.004.

Cambria, E., and B. White. 2014. "Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine* 9 (2):48–57. doi:10.1109/MCI.2014.2307227.

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., … & Kurzweil, R. (2018). Universal sentence encoder. arXiv preprint arXiv:1803.11175.

Chen, C. L. P., and C.-Y. Zhang. 2014. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Inf. Sci* 275:314–347. doi:10.1016/j.ins.2014.01.015.

Chen, Y., B. Perozzi, and S. Skiena. 2017. "Vector-based similarity measurements for historical figures. *Information Systems* 64:163–174. doi:10.1016/j.is.2016.07.001.

Cooper, R. 2017. "*Supply chain development for the lean enterprise: Interorganizational cost management.*Portland: Routledge Productivity Inc.

Cooper, R., and R. Slagmulder. 2004. "Interorganizational cost management and relational context. *Accounting, Organizations and Society* 29 (1):1–26. doi:10.1016/S0361-3682(03)00020-5.

Delhi, S. I. N. 2016. "Automotive revolution & perspective towards 2030. *Auto Tech Review* 5 (4):20–25. Springer Nature India Pvt Ltd, New Delhi.

Gloeggler, M. 2003. *Suchmaschinen im internet. funktionsweisen, ranking methoden, top positionen*. Berlin: Springer Verlag.

Goby, N., T. Brandt, S. Feuerriegel, and D. Neumann. 2016. "Business intelligence for business processes: The case of IT incident management. *European Conference on Information Systems, 24th European Confrence on Information Systems (ECIS 2016)At: Istanbul, Turkey*. DOI:10.13140/RG.2.1.2033.9604

Grigori, D., F. Casati, M. Castellanos, U. Dayal, M. Sayal, and M. C. Shan. 2004. "Business process intelligence.*Computers in Industry* 53 (3):321–343. doi:10.1016/j.compind.2003.10.007.

Gupta, V. 2019. "*Technology optimization and change management for successful digital supply chains*. Hershey: IGI Global.

Henke, N., J. Bughin, M. Chui, J. Manyika, T. Saleh, B. Wiseman, and G. Sethupathy. *The age of analytics: Competing in a data-driven world*. McKinsey Global Institute. vol. 4

Holzinger, A., and G. Pasi, "Human-computer interaction and knowledge discovery in complex, unstructured, big data", Third International Workshop HCI-KDD 2013, Springer Berlin Heidelberg, Berlin/Heidelberg, 2013.

Hong, L., and B. D. Davison, "Empirical study of topic modeling in twitter, Proceedings of the First Workshop on Social Media Analytics - SOMA '10, ACM Press, New York, 2010, 80–88.

Hung, R.-Y.-Y. 2006. "Business process management as competitive advantage: A review and empirical study. *Total Quality Management & Business Excellence* 17 (1):21–40. doi:10.1080/14783360500249836.

Hutchison, D., T. Kanade, and J. Kittler, "Computational linguistics and intelligent text processing, 14th International Conference CICLing 2013, Springer Berlin Heidelberg, Berlin/Heidelberg, 2013.

Kauffman, R. J., and H. Mohtadi, "Information technology in B2B e-procurement: Open vs. proprietary. 35th annual Hawaii international conference on system sciences, Hawaii, 2002, 2129–2138.

Kauffman, R. J., and H. Mohtadi. 2004. "Proprietary and open systems adoption in e-procurement: A risk-augmented transaction cost perspective. *Journal of Management Information Systems* 21 (1):137–166. doi:10.1080/07421222.2004.11045798.

Kearney, C., and S. Liu. 2013. "Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis* 33: 171-185.

Kleinhans, C., and K. Bräuning. 2015. *Automotive entwicklungsdienstleistung: zukunftsstandort deutschland; eine studie des verbands der automobilindustrie e. V. (VDA) in zusammenarbeit mit berylls strategy advisors GmbH"*. Berlin: Verband der Automobilindustrie.

Leopold, H., F. Pittke, and J. Mendling, "Towards measuring process model granularity via natural language analysis,Business Process Management Workshops, Springer International Publishing, Beijing, 2014, 417–429.

Leopold, H., J. Mendling, and A. Polyvyanyy. 2014. "Supporting process model validation through natural language generation. *IEEE Trans-actions on Software Engineering* 40 (8):818–840. doi:10.1109/TSE.2014.2327044.

Manning, C. D., and H. Schütze. 1999. "*Foundations of statistical natural language processing* Cambridge: MIT Press.

Manning, C. D., P. Raghavan, and H. Schuetze. 2018. "*Introduction to information retrieval.* Cambridge: Cambridge University Press.

Martens, D., F. Provost, J. Clark, and E. J. de Fortuny. 2016. "Mining massive fine-grained behavior data to improve predictive analytics. *MISQ* 40:869–888. doi:10.25300/MISQ/2016/40.4.04.

Mostafa, M. M. 2013. "More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications* 40 (10):4241–4251. doi:10.1016/j.eswa.2013.01.019.

Niederhoffer, V. 1971. "The analysis of world events and stock prices. *The Journal of Business* 44 (2):193–219. doi:10.1086/295352.

Olausson, D., T. Magnusson, and N. Lakemond. 2009. "Preserving the link between R&D and manufacturing: Exploring challenges related to vertical integration and product/process newness.*Journal of Purchasing and Supply Management* Elsevier Ltd. 15 (2):79–88. doi:10.1016/j.pursup.2008.12.004.

Oliver, J. R. 1996. "A machine-learning approach to automated negotiation and prospects for electronic commerce." *Journal of Management Information Systems* 13 (3):83–112. doi:10.1080/07421222.1996.11518135.

Orina, R. K., "Centralized purchasing strategies and organizational performance in the manufacturing industry", 2018.

Pang, B., and L. Lee. 2008. "Opinion mining and sentiment analysis, found trends. *Inf Retr* 2:1–135.

Paradkar, A., J. Knight, and P. Hansen. 2015. "Innovation in start-ups: Ideas filling the void or ideas devoid of resources and capabilities. *Technovation* 41-42:1–10. doi:10.1016/j.technovation.2015.03.004.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2013. Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

Piatetsky, G. 2019. "*CRISP-DM, still the top methodology for analytics, data mining, or data science projects*. KDnuggets.

Porter, M. E. 1985. "*Competitive advantage. Creating and sustaining superior performance*. New York: Free Press.

Pröllochs, N., and S. Feuerriegel. 2020. "Business analytics for strategic management: Identifying and assessing corporate challenges via topic modeling. *Information & Management 57(1)* 57 (1):103070. doi:10.1016/j.im.2018.05.003.

Pröllochs, N., S. Feuerriegel, and D. Neumann. 2016. "Negation scope detection in sentiment analysis: Decision support for news-driven trading". *Decision Support Systems* 88:67–75. doi:10.1016/j.dss.2016.05.009.

Qi, J., Z. Zhang, S. Jeon, and Y. Zhou. 2016. Mining customer requirements from online reviews: A product improvement perspective. *Inf. Manag* 53 (8):951–963. doi:10.1016/j.im.2016.06.002.

Ramage, D., D. Hall, R. Nallapati, and C. D. Manning. (2009, August). Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora. In Proceedings of the 2009 conference on empirical methods in natural language processing, 248-256, Singapore.

Rao, Y., H. Xie, J. Li, F. Jin, F. L. Wang, and Q. Li. 2016. "Social emotion classification of short text via topic-level maximum entropy model. *Information & Management* 53 (8):978–986. doi:10.1016/j.im.2016.04.005.

Ravi, K., and V. Ravi. 2015. "A survey on opinion mining and sentiment analysis: Tasks, approachesand applications." *Knowl.-Based Syst* 89:14–46. doi:10.1016/j.knosys.2015.06.015.

Reichuber, A. W. 2010. *Strategie und struktur in der automobilindustrie. Strategische Und Organisatorische programme zur handhabung automobilwirtschaftlicher herausforderungen*". Wiesbaden: Gabler Verlag/GWV Fachverlage GmbH.

Rus, V., N. Niraula, and R. Banjade, “similarity measures based on latent dirichlet allocation. 14th International Conference CICLing 2013, Springer Berlin Heidelberg, Berlin/ Heidelberg, 2013, 459–470.

Salama, D. M., and N. M. El-Gohary. 2016. “Semantic text classification for supporting automated compliance checking in construction. *Journal of Computing in Civil Engineering* 30(1):04014106.

Sarkar, D. 2016. “*Text analytics with python*. CA, Apress: Berkeley.

Sarkar, D., R. Bali, and T. Sharma. 2018. *Practical machine learning with python*. Berkeley, CA: Apress.

Shi, Z., G. Lee, and A. B. Whinston. 2016. “Toward a better measure of business proximity: Topic modeling for industry intelligence. *MIS Q* 40 (4):1035–1056. doi:10.25300/MISQ/2016/40.4.11.

Sparck Jones, K., and P. Willett. 1997. “*Readings in information retrieval*. San Francisco: Morgan Kaufmann.

Statista. 2010. Automobilzulieferer - Wertschöpfungsanteil der Automobilzulieferer am weltweiten Automobilbau in den Jahren 1985 bis 2015. https://de.statista.com/statistik/daten/studie/162996/umfrage/wertschoepfungsanteil-der-automobilzulieferer-am-automobilbau-weltweit/https://de.statista.com/statistik/daten/studie/162996/umfrage/wertschoepfungsanteil-der-automobilzulieferer-am-automobilbau-weltweit/ .

Subramaniyaswamy, V., V. Vijayakumar, R. Logesh, and V. Indragandhi. 2015. “Unstructured data analysis on big data using map reduce. *Procedia Computer Science* 50:456–465. doi:10.1016/j.procs.2015.04.015.

Tan, M. H., and W. L. Lee. 2015. “Evaluation and improvement of procurement process with data analytics. *International Journal of Advanced Computer Science and Applications* 6 (8):70. doi:10.14569/IJACSA.2015.060809.

Towne, W. B., C. P. Rosé, and J. D. Herbsleb. 2016. “Measuring similarity similarly: LDA and human perception.*ACM Transactions on Intelligent Systems and Technology* 8 (1):1–28. doi:10.1145/2890510.

Van der Aalst, W. M. P., K. M. Van Hee, J. M. Van der Werf, and M. Verdonk. 2010. Auditing 2.0: Using process mining to support tomorrow’s auditor. *Computer* 43 (3):90–93. doi:10.1109/MC.2010.61.

Walter, T. P., and A. Back, A text mining approach to evaluate submissions to crowdsourcing contests. In 2013 46th Hawaii International Conference on System Sciences (pp. 3109-3118). IEEE, Grand Wailea, Maui, Hawaii.

Wirth, J., and J. Hipp, “CRISP-DM: Towards a standard process model for data mining, 4th international conference on the practical applications of knowledge discovery and data mining, Springer, Wiesbaden, 2000, 29–39.

Zhu, T., and K. Li. 2012. “The similarity measure based on LDA for automatic summarization. *Procedia Engineering* 29:2944–2949. doi:10.1016/j.proeng.2012.01.419.