Heiko Hossfeld, Martin Wolfslast*

Text Classification in Organizational Research – A Hybrid Approach Combining Dictionary Content Analysis and Supervised Machine Learning Techniques**

Abstract

Big Data is an emerging field in organizational research as it provides new types of data, and technologies like digitization and web scraping allow to study huge amounts of data. Since large parts of digital data consist of unstructured text, text classification – assigning texts (or parts of texts) to predefined categories – is a central task. Text classification not only allows to identify relevant texts in a jumble of data but also to extract information from texts, such as sentiments, topics, and intentions. However, large amounts of textual data require the use of automated text mining methods, which is mostly uncharted territory in organizational research. We, therefore, outline and discuss the two existing approaches to text classification, one originating from social science (dictionary content analysis) the other from computer science (supervised machine learning). Since both approaches have advantages and disadvantages, we combine ideas from both to develop a hybrid approach that reduces existing issues and requires significantly less knowledge in programming and computer science than supervised machine learning. To illustrate our approach, we develop a classifier that identifies critical media coverage of organizational actions.

Keywords: text classification, text mining, big data, organizational legitimacy
(JEL: C11, C18, M14)

Introduction

Technological advances in digitization are causing fundamental changes in society, business, and work, frequently discussed under buzzwords like “digital transformation” (Andal-Ancion et al., 2003) and “Industry 4.0” (Piccarozzi et al., 2018). A by-product of this development is new data available to organizational research, for example, in the form of online data (e.g., social media data) or mobile data (e.g., geographical location) (Sheng et al., 2017). In addition, the amount of available data has increased rapidly due to digitization. This includes data that was previously

* Dr. Heiko Hossfeld (corresponding author), University of Labour, Frankfurt, Germany, Email: heiko.hossfeld@university-of-labour.de
Dr. Martin Wolfslast, zeb.rolfes.schierenbeck.associates gmbh, Muenster, Germany. Email: martin.wolfslast@zeb.de.

** Date submitted: August 17, 2020
Date accepted after double-blind review: March 27, 2021.

DOI: 10.5771/0935-9915-2022-1-59

only available in printed form, like annual reports and newspaper articles. A large part of this "Big Data" is unstructured text (Cogburn & Hine, 2017; Kobayashi et al., 2018b).

So far, however, analyzing textual data in the context of Big Data (i.e., textual data characterized in particular by large volume and high velocity) has played a relatively minor role in organizational research (Kobayashi et al., 2018b; Sheng et al., 2017). This is problematic because processing large text corpora or large text collections is a prerequisite for numerous Big Data applications (e.g., real-time social media analysis). Moreover, this is quite surprising because handling textual data is not uncommon in management and organization studies (Kabanoff, 1997; Pollach, 2012), especially since this kind of data is mostly gathered non-obtrusively (George et al., 2016) and does not suffer from hindsight bias (Platanou et al., 2018). Textual data can be used for a wide range of research purposes by analysing not only the manifest content of the text (like word occurrences or frequencies) but also the latent content embodied in the text (Duriau et al., 2007), for example political positions (Laver & Garry, 2000), managerial discourses (Vaara & Tienari, 2002), and the rhetoric of management (Hossfeld, 2018). Furthermore, textual data can be used both for exploratory purposes (e.g., theory development) and theory testing (George et al., 2016; Kobayashi et al., 2018b). In the latter case, textual data is often combined with quantitative data like stock prices (e.g., Ahmad et al., 2016).

At first sight, digitization merely seems to extend the existing body of data for analyzing the communication of organizations (e.g., reports, social media, patents), in organizations (e.g., social intranet), to organizations (e.g., application documents), and about organizations (e.g., news media, Twitter, review sites). However, the greater impact on organizational research stems from the fact that the technologies associated with digitization (e.g., web scraping[1]) have made it much easier to collect large numbers of documents (e.g., annual reports or media coverage) – and to do so in real time if necessary. Unless the goal is to qualitatively analyze small samples, methods are needed that can handle large text corpora or text collections. Yet, although "computer-aided text analyses have gained a lot of attention recently" (Bannier et al., 2019, p. 79), organizational research has no established research tradition of automated text analysis, albeit such methods already exist. These techniques are summarized under the term text mining (Bonfiglioli & Nanni, 2015). Text mining allows for rapid coding of very large amounts of text data. This has three main advantages: First, larger sample sizes can be studied in this way. This is particularly important for theory testing in organizational research, for example, to investigate the effect of media coverage on stock prices (e.g., Ahmad et al., 2016). The other two benefits are more relevant to organizational practice: often, random

[1] With standalone software like Octoparse or Zyte and Python packages like Scrapy or MechanicalSoup, automatically extracting large sets of text data from the web has become much easier for non-programmers.

samples are not sufficient, but complete collections of texts are to be studied (e.g., when analyzing patient records or personnel data). In addition, automatic procedures allow real-time coding, which is necessary for social media analyses, for example.

A very common and important objective in text mining is text classification, i.e., assigning text – whole documents or parts of a text (e.g., sentences or paragraphs) – to predefined categories (Harish et al., 2010; Kobayashi et al., 2018a). In some cases, text classification is merely used to identify and select texts for further analysis, for example, thematically relevant articles from a database of scientific journals (e.g., Bansal & Gao, 2006) or from a larger corpus of media coverage (e.g., Strycharz et al., 2018). Often, however, text classification produces the key variables of the analysis. This is, for example, the case with sentiment analysis, e.g., in studies that survey the (positive or negative) tone of media coverage (Ahmad et al., 2016) or which identify disgruntled employees through their communication (Holton, 2009). Other forms of text classification include topic classification (e.g., classifiying job tasks from nursing vacancies, Kobayashi et al., 2018a) and intent classification. The latter can be used, for example, to categorize customer intentions, such as questions or suggestions (Pérez-Vera et al., 2017), or to detect racism in social media posts (Agarwal & Sureka, 2016).

There are two different approaches to text classification which have been developed independently of each other. Dictionary content analysis classifies texts on the basis of wordlists. As a further development of conventional content analysis, it originates from social science research and is also occasionally used in organizational research (especially in sentiment analysis). Supervised Machine Learning, on the other hand, originates from computer science and is, accordingly, characterized by a higher degree of mechanization. So far, it has been largely ignored in organizational research, partially because building computer algorithms requires specific expertise.

In this paper, we discuss both approaches and argue that neither is superior to the other per se since both have shortcomings – more researcher subjectivity and effort in one case, less transparency, and reproducibility in the other. Therefore, we propose a hybrid approach to text classification which combines the advantages of both methods and can be performed without in-depth computer science knowledge. By doing so, we contribute to an emerging field in organizational research. To demonstrate the applicability of our approach, we used it to automatically identify press articles reporting critically on the actions of the thirty largest German companies.

Approaches to Text Classification

It is not without reason that content analysis is very popular in research in the social sciences and in organizational science. It has always stood at the "intersection of the qualitative and quantitative traditions" (Duriau et al., 2007, p. 5) and, ideally, combines the advantages of both: on the one hand, the systematic coding of text allows non-numerical data to be included in quantitative analyses (George et al., 2016). On the other hand, manual coding helps to ensure that researchers correctly grasp the complexity of natural language (e.g., metaphors or sarcasm, cf. Guo et al., 2016) and take this into account when classifying text. Yet, content analysis was not originally designed to handle large bodies of text (Holsti, 1969; Lewis et al., 2013), and Big Data explicitly "refers to datasets that are too big for humans to code" (Guo et al., 2016, p. 333). Large text data can only be coded and classified automatically using algorithms.

Text mining research distinguishes two general approaches to text classification: knowledge engineering is based on logical classifiers, which means that the researchers manually define a set of logical rules about how to classify texts under given categories. The machine learning approach, in contrast, builds an automatic text classifier – typically either based on geometrical or probabilistic algorithms – by learning from a set of pre-classified reference texts (Kobayashi et al., 2018a; Yehia et al., 2016). Along with these two directions, two different methods of text classification have been developed: dictionary content analysis and supervised machine learning.

Dictionary Content Analysis (DCA)

Since manual coding is time-consuming, expensive, and error-prone (Xie & Xing, 2018), methods and software solutions were developed as early as the 1960s to support content analysis with the aid of computers (Pollach, 2012). Today, computer-aided content analysis or CATA (computer-aided text analysis) is the most popular set of methods for text analysis in organizational studies (Kobayashi et al., 2018b). In some cases, CATA is used primarily to support researchers in the manual coding of text. However, CATA has also been used from the very beginning to assign text units automatically to a coding scheme, aiming to replace the coding by experts (Laver & Garry, 2000; Pollach, 2012). This has led to a subfield of CATA called the dictionary-based approach (Bannier et al., 2019; Pollach, 2012) or dictionary content analysis (Kothari et al., 2009).

The basic idea of dictionary content analysis is very simple: First, the researcher defines one or more dictionaries, i.e., lists of words or phrases that are systematically associated with categories like teamwork or leadership. Then the computer counts the number of terms per category in each text document (Laver & Garry, 2000; Riffe et al., 2019) – using either a specialized software (like NVivo or MaxQDA) or a general programming language (like Python or R). This approach is based on a simplifying assumption since documents are represented as "bags of words," where a text is reduced to the occurrences and frequencies of the words used within while word order and syntactic relations between words are not taken into account (Pollach, 2012). Although or maybe because this method ignores – and thus reduces – the complexity of natural language, it has a long and successful history in the social sciences (e.g., Bannier et al., 2019; Pollach, 2012; Riffe et al., 2019). Dictionary content analysis is also used in organizational and management research, especially for sentiment analysis (e.g., Castelló et al., 2016; Loughran & McDonald, 2015; Pollach, 2012).

Sentiment analysis is used to measure the tone of texts based on the vocabulary used, for example, when analyzing how companies, their practices, or products are discussed in the news (e.g., Ahmad et al., 2016) or social media (Ghiassi et al., 2013) or when studying how employees talk about their work and employer (e.g., van Zoonen et al., 2016). For dictionary-based sentiment analysis, a whole range of software solutions with integrated general dictionaries has been developed in recent decades, e.g., LIWC (Pennebaker et al., 2015) or Senti-WordNet (Baccianella et al., 2010). These dictionaries either make a very basic distinction between terms associated with a positive emotion (e.g., "exciting", "fair", "safe") or a negative one (e.g., "rude", "fear", "poor") or make a more differentiated distinction between emotions like anger ("cruel", "brutal", "stupid") and anxiety ("afraid", "doubt", "panic").

To classify texts, for example, as either positive or negative, dictionary-based approaches use logical classifiers. Typically, sentiment analyses either use the relative frequencies of words from a dictionary (for example, the percentage of positive words) or they put the number of positive terms in relation to the number of negative terms (e.g., Bae & Lee, 2012; Caserio et al., 2019). The researchers then either define thresholds (e.g., at what point a text is classified as negative) or use the relative frequencies as the basis for a metric classification, for example, as an indicator of sentiment intensity. In this case it is, e.g., assumed that the more terms from the category "anger" are used, the "angrier" the text is. For other tasks, e.g., in the field of topic classification, absolute frequencies are often used to define logical classifiers. Here, texts are often already assigned to a certain category if they use at least one term from a dictionary, for example, to classify types of qualifications mentioned in job descriptions (Park et al., 2009).

Although choosing the most accurate classifier is an important task in dictionary content analysis, the quality of the dictionary is crucial. For this reason, a wide range of standard dictionaries has been developed over time that are useful for many social science contexts (Garrad, 2003). In addition to sentiments, these dictionaries cover textual features like the use of personal pronouns and negations, as well as more abstract categories like values, orders of worth, and policy positions (for an overview, see Humphreys & Wang, 2018). Since creating and validating dictionaries requires a lot of time and effort, organizational researchers (among others) often prefer to use established standard dictionaries (Pollach, 2012). However, there are no ready-made dictionaries for every research topic. This is one reason why organizational researchers also use self-constructed dictionaries, for example, to study CSR disclosures (Pencle & Mălăescu, 2016) or the language of downsizing (Hossfeld, 2013). The other reason is that language is alive and constantly changing. Every domain (e.g., business, journalism, science, social classes) has its specific vocabulary and words often vary in their meaning across domains. For example, Loughran and McDonald (2016) show that many terms classified as negative (e.g., cost, liability, gross) or positive (e.g., trust) in established standard dictionaries have a different meaning in financial texts. Therefore, a dictionary built from one domain (or corpus) may not be transferable to other domains. This is why domain-specific dictionaries are often preferable to standard dictionaries (Loughran & McDonald, 2015; Pollach, 2012).

To build a good dictionary for text classification, it is necessary to identify highly selective terms, i.e., words that are mainly used in only one category of text. This is done either deductively from theory or inductively from the text corpus of interest (or by combining both methods, Pollach, 2012). In both cases, dictionaries are typically developed manually by researchers (but with the assistance of computers) (Guo et al., 2016). This semi-automatic approach to text classification has the advantage that researchers have a high degree of control over the dictionary development process, but this is also its biggest weakness since the approach relies heavily on the experiential worlds and vocabularies of the researchers.

This is especially true for the deductive method as it completely ignores domain-specific features of the language, i.e., the vocabulary of the corpus' authors. The inductive method tries to solve this problem by searching the vocabulary of the corpus for relevant terms in a time-consuming manual coding process. However, researcher subjectivity makes these processes error-prone (Loughran & McDonald, 2016) as it can cause relevant terms to be overlooked or researchers to consider terms important that are actually irrelevant.

Supervised Machine Learning (SML) Approach to Text Classification

In contrast to dictionary content analysis, supervised machine learning is a (mostly) automatic method of text classification. Machine Learning, in general, is a wide field in computer science that deals with the study of methods for pattern recognition in data, including (but not limited to) unstructured text (Aggarwal, 2018). When it comes to text mining, the most important categories of machine learning are supervised and unsupervised learning (Prüfer & Prüfer, 2018). Both of them provide broad application possibilities for organizational research but have hardly been utilized by organizational researchers to date (Tonidandel et al., 2018; Wenzel & Van Quaquebeke, 2018). Unsupervised machine learning, being a tool of exploratory data analysis, is not relevant for our purpose as it is a learning algorithm that draws inferences from unlabeled data (Prüfer & Prüfer, 2018). The two most popular approaches here are cluster analysis, which is used to group whole texts according to shared characteristics, and topic modelling, which is used to identify topics in texts by examining the pattern of word frequencies (Kobayashi et al., 2018b).

In contrast, supervised machine learning is a method of text classification as it is used to code large numbers of text documents into predefined categories. Although there are several different approaches, supervised machine learning is always based on a set of texts already pre-coded for the categories of interest (van Zoonen & Toni, 2016), for example, texts that share the same policy position (Laver & Garry, 2000), job task (Kobayashi et al., 2018a), or sentiment (Huang et al., 2014). This set of documents, which is usually manually coded by experts, serves as training data: The machine learning algorithm analyzes these reference texts and produces an inferred function that is used for classifying novel data (Prüfer & Prüfer, 2018), for example, to predict legal court decisions (Martin et al., 2004). The most common approaches to supervised machine learning in text mining use either geometric or probabilistic algorithms. While geometric algorithms (e.g., K-nearest neighbours and support vector machines) are based on the assumption that texts can be represented as points in a multi-dimensional space, probabilistic algorithms are grounded in probability theory (Kobayashi et al., 2018a).

Among the probabilistic algorithms, naïve Bayes classifiers are very popular (Ikonomakis et al., 2005; Kobayashi et al., 2018a; Loughran & McDonald, 2015). Like dictionary content analysis (and most of the other approaches in supervised machine learning, Aggarwal, 2018), this method treats texts as bags of words. For a probabilistic algorithm, this translates into the simplifying assumption that text features (here: words) in a dataset are mutually independent (hence "naïve" Bayes). Since each text is represented as a vector of word counts (or occurrences), the words used in a text serve as predictors for determining the probability of that text belonging to a certain class. In the multinomial naïve Bayes model, for example, the relative word frequencies in the reference texts are used to determine P(W | C), which is the probability of word w in class c. Then an algorithm computes P(C | T), i.e., the probability that the text t belongs to class c, for the word vector of each novel text and assigns the text to the class with the highest probability (for a detailed description of the method see Aggarwal, 2018).

Despite its simplifying assumptions – since words do not always occur independently of each other (e.g., “poison pill”) –, the naïve Bayes classifier is rather effective and robust (Evans et al., 2007; Feldman & Sanger, 2007; Ikonomakis et al., 2005). Although machine learning algorithms are still used relatively rarely in the social sciences, a method for text classification has become established in political science that is based on assumptions very similar to those of naïve Bayesian algorithms (Evans et al., 2007): the wordscore method which is used to estimate policy positions in texts (Laver et al., 2003). Word scores are calculated based on the relative frequency of each word in each reference text and considering the class assigned to the document. These word scores are then used to classify new texts according to their vocabulary. In both cases – naïve Bayes algorithm and wordscore method –, each term used in the training data set helps to determine which class a text belongs and the more selective a term is, the greater its contribution to text classification.

Combining DCA and SML in a Hybrid Approach to Classifying Texts Automatically (HACTA)

When comparing both methods (Table 1), it is noticeable that DCA and SML are based on the same assumption: classes of texts differ according to the vocabulary used; thus, words are used as features for a classification algorithm. Although both methods differ regarding how a classifier is developed from this bag-of-words assumption, both approaches have proven that they are capable of producing effective classifiers – at least when comparing SML to DCA with domain-specific dictionaries, since the applicability of standard dictionaries depends on the object of investigation (Bannier et al., 2019). Both methods also classify fully automatically, so both are well suited for handling large data sets.

Table 1: Comparison Between DCA and SML

	DCA	**SML**
Key assumption	Bag of words	Bag of words (usually)
Potential for accurate classifiers	Standard dictionaries: low-high Domain-specific dictionaries: high	High
Degree of automation of the classification process	High	High
Degree of automation of the classifier development	Low	High
Problem of researcher subjectivity	High	Low – medium
Effort	Standard dictionaries: low Self-constructed dictionaries: very high	Medium But: expertise in computer science required
Transparency and reproducibility	High	Low

In some other aspects, however, the two differ considerably: DCA is easy to use and has high transparency and reproducibility (as the dictionaries can be viewed and used by other researchers). However, feature selection, i.e., building a valid dictionary, is a problem unless an established dictionary can be used. Developing dictionaries manually is time consuming and dependent on human coders to find highly selective terms. This makes the method susceptible to issues caused by researcher subjectivity since researchers are not necessarily able to assess which terms predominantly occur in only one text category. Researchers might, e.g., overlook highly selective terms or assume that terms are selective when they are not. These issues cannot be solved completely, even by involving multiple researchers.

SML has a significant advantage here because feature selection is much faster, more objective and based on the domain-specific language. The naïve Bayes algorithm determines the explanatory contribution of each individual word based on statistical probabilities. This reduces the subjectivity problem, although human error can also occur in the classification of reference texts as this is usually done manually (which is why intercoder reliability must be ensured, Aggarwal, 2018). However, the fully automated approach of SML has other issues: Building computer algorithms requires specific expertise which organizational researchers still rarely have (Kobayashi et al., 2018b). This may change, though, once text mining becomes more established in organizational research. Nevertheless, several other issues remain. The procedure is not very transparent for other researchers and, therefore, difficult to replicate because the training data set is usually not publicly available and many rules and filters used in machine learning are poorly documented (Bannier et al., 2019; Feldman & Sanger, 2007; Loughran & McDonald, 2016). Furthermore, the use of this method also means a loss of transparency and control for the researchers if they rely entirely on the computer algorithm (Stulpe & Lemke, 2016). This is, for example, reflected in a problem related to overfitting, i.e., when an algorithm performs well on the training data but poorly on new data (Kobayashi et al., 2018b). This happens, e.g., when words enter the algorithm that are only selective for the training data set but not for other data. One way to solve this problem would be to validate the features with the use of qualitative methods, for example, by conducting a (key)word in context (KWIC, Bernard et al., 2016) analysis.

Thus, a good method for automatic text classification cannot rely on statistical algorithms alone but also on qualitative expertise – both in the coding of reference texts and in feature selection. Therefore, we argue that it makes sense to combine DCA and SML since both approaches have their own specific advantages and disadvantages. In this way, we can minimize the disadvantages while retaining the advantages. Our hybrid approach to classifying texts automatically (HACTA) does exactly that. We adopt the basic idea of DCA, which is easy to use and has high transparency and reproducibility, i.e., we classify texts via a logical classifier using dictionaries of highly selective terms. Contrary to the conventional method, though, dictionaries are not generated (primarily) manually, but in a two-step process: First, we make use of the advantages of SML by using a statistical algorithm to identify highly selective terms for a training data set. The resulting raw dictionary is then qualitatively analysed, especially to avoid (or reduce) overfitting issues. The process of HACTA is outlined in Table 2.

Table 2: Process of HACTA

Name of step	Tasks	Aim
1) Text preprocessing	■ Data scraping ■ Generating document-term matrix ■ Data cleaning (and dimension reduction)	Two data sets, one for qualitative analysis (complete documents) one for quantitative analyses (document-term matrix)
2) Training data preparation	■ Developing coding guidelines ■ Drawing a random sample of documents ■ Manual coding of the sample	Training data set
3) Quantitative analysis	Identifying selective terms	Raw dictionary
4) Qualitative analysis	Validating raw dictionary	Final dictionary
5) Classifier development	Developing logical rules for classification	Probabilistic or geometric algorithm

Using HACTA to Classify Media Coverage

In this section, we will illustrate our hybrid approach with a concrete example from organizational legitimacy research. Our aim is to build a classifier that automatically identifies press articles reporting critically on an organization’s actions (i.e., it distinguishes between critical and non-critical articles). A critical article is characterized by doubting the legitimacy of organizational action, either directly by the journalist or indirectly by reporting critical voices. If this happens on a larger scale, the "licence to operate" (Scherer & Palazzo, 2011, p. 914) and with it critical resources may be withdrawn (Pfeffer & Gerald, 1978). Studies show that negative coverage, regardless of its veracity, can have negative consequences for companies (Durand & Vergne, 2015).

The subjects of our examination were the 30 biggest German companies (DAX 30) over a period of ten years (2004–2013). The data stem from online media coverage by two influential German weekly news publications (Spiegel Online and Zeit Online). In the period we examined, the business sections of the two news sites published over 62,000 articles in total, making a purely qualitative text analysis nearly impossible.

Step 1: Text Preprocessing

Every text mining application starts with text preprocessing, which is a series of steps used to prepare and clean raw text data for further processing (e.g., Aggarwal, 2018). Several software solutions exist for this purpose. R and Python are especially worth mentioning as many free text mining packages are available for both. First, the relevant data (e.g., article text and date) are extracted from the document files using data scraping and are aggregated into a single data set (rows = documents, columns = a single string variable for the text as well as variables for date and source). A copy of this data set is stored for the qualitative analysis in step 4.

A bag of words representation of the data is then created for quantitative analysis: the document-term matrix (DTM) describes how often each term occurs in a collection of documents (rows: documents, columns: terms). Thus, each article is a vector of terms existing in the dataset. However, to ensure that the DTM contains only relevant text elements, data cleaning, and dimensionality reduction must be performed first. The text data is cleaned by transforming all words to lower case and by removing punctuation, numbers, and excess spaces. Document-term matrices usually have many variables (terms), so it is advisable to reduce this number for analysis (Kobayashi et al., 2018a). For this purpose, stop words are removed. These are common words such as "a, the, of" that do not carry much discriminative content. In addition, all words are reduced to their common root to merge, for example, the terms "scandal" and "scandals" into one variable. This is usually either done by stemming or by the more sophisticated lemmatization (see in detail Aggarwal, 2018).

For many applications, the text preprocessing would be finished now, but here, a preliminary text classification was required. Since the goal of our analysis is to identify articles that problematize the actions of specific companies, the articles that deal with the DAX 30 companies had to be identified first. This is usually done automatically using the following rule: If an article includes the name of the company in the header, it is classified as an article about that company (e.g., Bednar, 2012). However, this will also include articles that name several companies at the beginning or are really about a different company but still mention the name of the focal organization in the header. We, therefore, extended the logical classifier to consider the length of the text: the longer the article, the more often the company name must be mentioned. This was done based on a qualitative analysis of 200 articles. In total, about 10 percent of the articles examined (6,492) are about one of the DAX30 companies.

Step 2: Training Data Preparation

Since a pre-coded training data set is required for the quantitative analysis, a random sample of articles is drawn first. In our case, there were even two: A sample of 600 articles from the total corpus and 500 articles from the subcorpus of 6,492 company articles. Using coding guidelines, two researchers identified articles that reported critically on an organization’s actions (i.e., they classified the documents into the categories “critical” or “neutral”). Since it is only possible to identify highly selective terms if the text categories are clearly distinguishable from one another, only those texts that could be clearly assigned to one of the two categories were included in the quantitative analysis. Ambiguous texts, especially articles that were coded differently by the two researchers, were therefore excluded.

Step 3: Quantitative Analysis

In principle, there are several possible quantitative methods for identifying highly selective terms on the basis of a DTM; for example, naïve Bayes or document scaling approaches (like the aforementioned wordscore approach or the newer Latent Semantic Scaling method, cf. Watanabe, 2020) could be adapted accordingly. We chose another variant – albeit related to naïve Bayes – which is both user-friendly and intuitively interpretable: We calculated selective terms by comparing absolute and expected values.

First, we aggregated the sample´s DTM into a contingency table by grouping the reference texts into two subcorpora: articles that criticize organizational behaviour and neutral articles (Table 3). This table shows the absolute frequency (observed value) of each term in each class of texts as well as the marginal totals. Then, similar to a chi-squared test, we calculated the expected values under mutual independence ($e_{1K} = o_{.K} * o_{1.} / N$), i.e., the absolute term frequencies to be expected if the use of words was independent of the type of text. A comparison of both values provides a decision-making basis. To ensure that only highly selective terms are included in our raw dictionary, we included only those terms that were used twice as often in critical texts than would be expected given statistical independence ($o_{1k}/e_{1k} >= 2$).

Table 3: Contingency Table of Classes and Terms

	Terms x1 to xk				
Classes y	x_1	x_2	...	x_k	∑
Critical (y_1)	o_{11}	o_{12}	...	o_{1k}	$o_{1.}$
Non-critical (y_2)	o_{21}	o_{22}	...	o_{2k}	$o_{2.}$
∑	$o_{.1}$	$o_{.2}$	...	$o_{.k}$	N

Step 4: Qualitative Analysis and Feature Selection

The next step is to perform a KWIC analysis of the terms in the raw dictionary. The primary objective of this qualitative analysis is to ensure that the final dictionary does not contain terms that are highly selective only for the training data set. Therefore, to avoid overfitting, context-specific terms must be excluded.[2] In our data, this mainly happened when a training data set included several articles dealing

[2] In principle, the dictionary could also be expanded in this step. For example, collocation analysis or unsupervised machine learning methods could be used to identify terms (e.g., latent semantic scaling, cf. Watanabe, 2016) that occur frequently with the highly selective terms found. We have refrained from doing so here, as there is a high risk of identifying terms that are not selective for the actual text classification task.

with the same corporate scandal. In that case, particularly terms associated with the respective companies (e.g., the name of the industry, the company, or people involved) were identified as highly selective.

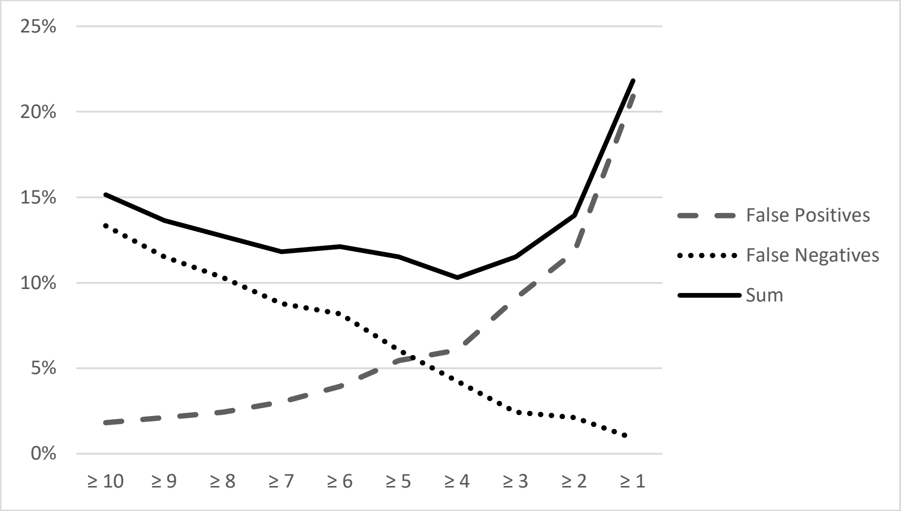
The second objective of our study was to build subcategories to be able to identify the kind of organizational behaviour criticized. For this purpose, we categorized the terms of the dictionary with a combination of correlation and KWIC analysis.

Step 5: Classifier Development

After the dictionary has been completed, a logical classifier must be defined. The researchers must first determine whether only the number of dictionary terms is relevant or also other text features, e.g., the text length or the position of dictionary terms (such as the header). Since it is irrelevant for our study whether an entire article critically reports on organizational behaviour or only parts of it, we developed a classifier that exclusively considers the number of dictionary terms.

We decided on the exact number with the help of the training data set: We took the number of dictionary terms at which the sum of false positives (articles are classified as critical when they are not) and false negatives (articles are not classified as critical when they are) is lowest as a threshold. This is the case for 4 terms, as shown in Figure 1. Thus, in the automatic classification, those articles that contain at least 4 terms of our dictionary were classified as "critical."

Figure 1: Percentage of False Positives and False Negatives in the Reference Texts, Depending on the Number of Dictionary Terms Used As Threshold



Results and Validation

The final dictionary is the result of an iterative process. First, we developed a first version of the dictionary based on two manually coded training data sets (with 600 and 500 articles, respectively). In order to extend this dictionary by additional terms and make it independent of the limited sample size, we then ran the quantitative analysis on two larger subcorpora; all 6,492 company articles and a sample of 10,000 other articles from the original corpus. We classified these texts into two groups: articles that do not contain any terms from the current raw dictionary (neutral) and articles that contain at least five terms from the dictionary (critical). Applying our method to the (manually coded) training data led to satisfactory results. In the course of the iterative process, the correlation between the category and the number of dictionary terms increased from 0.65 to 0.76. Our final classifier correctly assigns 90 percent of the reference texts. To assess whether the accuracy of 90 percent is satisfactory, our results should be compared with the results of other classifiers that measure critical coverage. Unfortunately, to the authors' knowledge, no such classifiers exist. However, we can compare our results with classifiers that measure something similar in content: For example, NLP classifiers used for sentiment analysis achieve comparable results (according to a meta study of Heitmann et al., 2020, the median accuracy across all data sets is 89 percent). Another example: the accuracy of existing SML methods for classifying political positions is usually below 80 % (Hausladen et al., 2020). Therefore, the accuracy of our classifier can be cautiously estimated as relatively high.

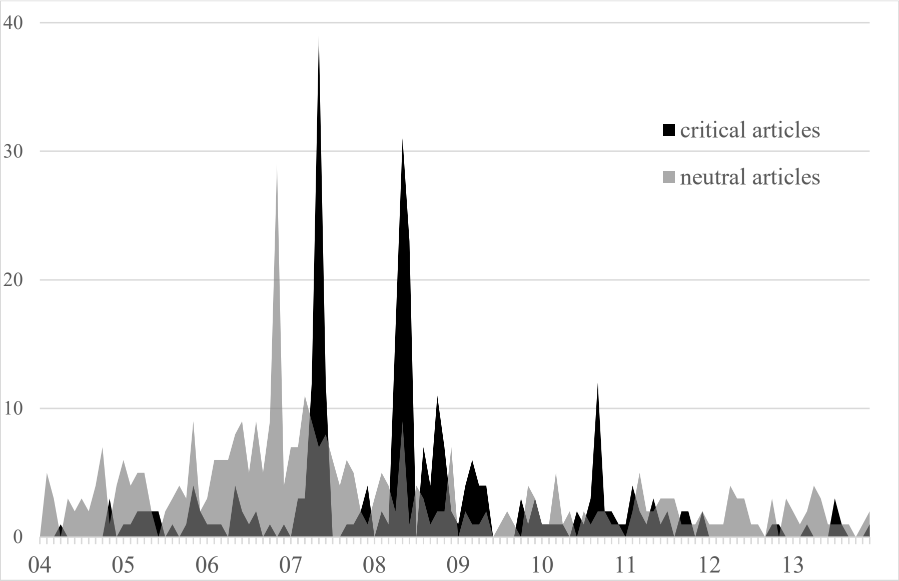
Table 4: Subcategories of the Dictionary

Category	**Number of terms**	**Examples of dictionary terms** **(translated from German)**
General categories		
Legitimacy	130	affair, misconduct, manipulation, shady
Legality	91	charges, penalty, illegal, witness, prosecutor
Type of organizational action		
Downsizing	13	downsizing, mass layoffs
Labour conflicts	20	strike, wage cut, labour dispute
Data abuse	11	data abuse, spy on
Fraud	3	embezzle, on company expenses
Mismanagement	11	mismanagement, bust, ruin
Cartel violation	5	price fixing, federal cartel office
Corruption	8	bribe, slush money, corrupt
Tax offence	9	money laundering, financial supervision
Consumer-related	5	customer complaints, consumer protection

The final dictionary contains a total of 286 terms or rather lexemes due to lemmatization. One aim of the qualitative analysis was the possibility to draw conclusions about the type of organizational behaviour criticized. It turned out that only some of the terms could be clearly assigned to a certain type of organizational action, while most of the identified terms were rather non-specific. These terms were assigned to two further categories: terms that “only” question the legitimacy of organizational action and those indicating that the organizational behaviour also has legal consequences (Table 4).

The dictionary holds some surprises. For one thing, some terms are not included that would probably have been included in a deductive procedure. This is the case, for example, with the term "crisis," which was not proven to be selective in statistical analysis. On the other hand, the dictionary contains terms that are surprising because they appear value-neutral, such as downsizing or working conditions. Not every case of downsizing is problematized, of course, but, at least in our training data set, the term is used more often in critical articles than in neutral ones. This does not necessarily have to be the case in other text corpora, which is why it can make sense to specifically exclude categories of terms in further analyses.

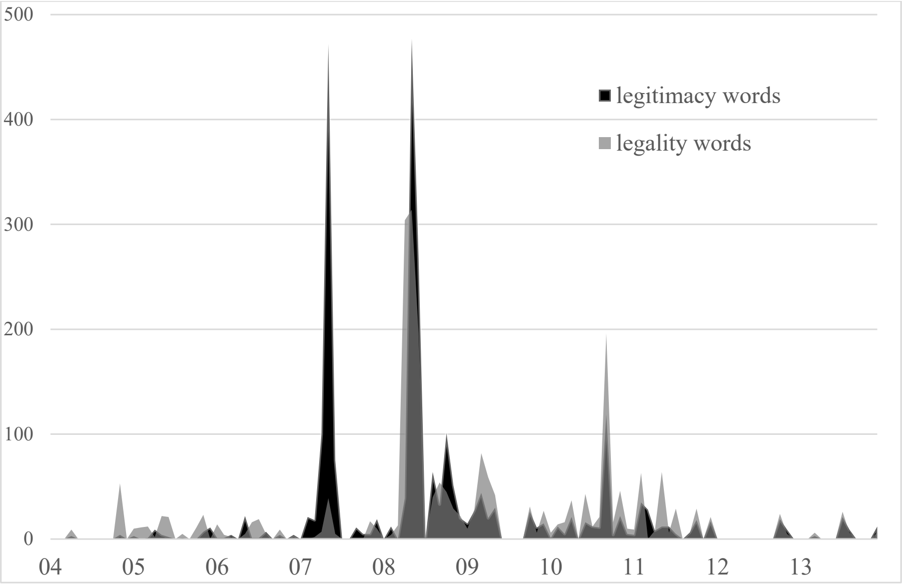
Figure 2: Number of Critical and Neutral Articles Over Time (Deutsche Telekom)



By applying the final dictionary to the entire corpus, the media coverage of the DAX 30 companies can be studied. Here, it is not so much the individual articles that interest us, but rather agglomerations of critical articles, since these point to concrete organizational scandals. We show this using the example of Deutsche

Telekom. Looking at the frequency of critical articles over time, we see two major spikes: one in 2008 the other in 2009 (Figure 2). To obtain better insight into what kind of behaviour is subject to criticism, we can look at the number of dictionary terms used per sub-category. First, we can see that, in 2007, almost exclusively legitimacy words were used, whereas in 2008, legality words were also used on a larger scale (Figure 3). The first event (or "scandal") apparently had no legal implications, whereas the second one did. We then use the same approach to look at the sub-categories that provide information about the nature of organizational behaviour. Here we find almost exclusively terms of the category "labour conflicts" in 2007 and, in 2008, terms of the category "data abuse." A subsequent qualitative analysis of the critical texts in 2007 and 2008 confirms the impression of the quantitative analysis: In 2007, Deutsche Telekom was criticized for outsourcing 50,000 jobs, which was legal but met with considerable resistance from the trade union. 2008, on the other hand, was dominated by the "eavesdropping controversy" (spying on several stakeholders), which led to investigations by the public prosecutor's office.

Figure 3: Number of Legitimacy and Legality Words Over Time (Deutsche Telekom)



To validate our method and to identify organizational scandals for all companies, we analysed all 1,953 critical articles manually. In this qualitative analysis, we identified 42 scandals (with at least 5 critical texts). Although we did not systematically recode the texts classified as critical by our classifier, only a very small number of false positives was identified, i.e., texts that our classifier had wrongly classified as critical. We did not identify any organizational scandals that had previously been overlooked. Due to its sheer size, it is difficult to say anything about false negatives within the entire text corpus. However, we tried to identify at least those articles that deal with one of the 42 scandals but were overlooked by our classifier. This number is negligible and mainly concerns articles just below the threshold of four dictionary terms.

We also performed external validation using data from Thomson Reuters. The company employs 150 data analysts who analyse media coverage of companies to identify media controversies (Thomson Reuters, 2017). Although Thomson Reuters does not clarify sufficiently how they operationalize a media controversy, their data is comparable with ours. Therefore, we conducted a correlation analysis which revealed a strong positive relationship between the number of controversies and the number of critical articles ($r_P = 0.55$) as well as the number of scandals ($r_P = 0.69$). This suggests that both methods measure something very similar.

Conclusion

In the age of digitization and big data, text mining is an emerging field in organizational research. Text classification tasks are becoming increasingly important in this context since text classification enables organizations to effectively use available information (Kobayashi et al., 2018a), e.g., when screening social media data. Furthermore, text classification is necessary for theory testing research, which benefits from the fact that digitization allows larger sample sizes. Despite text classification's great potential, automatic methods are still rarely used in organizational research. This paper aims to change that, not only by discussing and comparing the two existing approaches (dictionary content analysis and supervised machine learning) but also by presenting a hybrid approach to classifying texts automatically (HACTA) and illustrating its application with a concrete example.

A comparison of the established approaches shows that, while both are capable of developing accurate classifiers, both have distinct shortcomings: DCA is more prone to the problem of researcher subjectivity, and the development of dictionaries is very time consuming, whereas SML lacks transparency and reproducibility and requires expertise most organizational researchers currently do not have. The hybrid approach we propose reduces these problems by combining ideas from both approaches: As in DCA, text classification is based on dictionaries, but the process of dictionary development builds on the basic idea of SML since highly selective terms are identified statistically on the basis of a training data set. A subsequent qualitative analysis of the raw dictionary completes the process of this blended approach (Lewis et al., 2013). The result is a method that has the transparency and reproducibility of DCA (since dictionaries can be viewed and used by other researchers) and, at the same time, is as objective and time efficient as SML.

While HACTA is much more accessible than SML, it still requires some text mining expertise, namely some knowledge of text preprocessing (for a good introduction, see Aggarwal, 2018; Ignatow & Mihalcea, 2016) as well as basic programming skills in Python or R. In the project presented here we used R, an overview of the packages and functions used can be found in appendix 1. The fact is, if organizational research wants to benefit from the new data that have emerged in the course of digitization, organizational researchers will have to acquire some knowledge of text mining and machine learning. However, this should not be a reason to leave text mining to computer scientists. Knowledge in empirical social research, as well as subject matter expertise, are required throughout the text classification process. This is explicitly taken into account in HACTA, as was made clear in our example: On the one hand, expertise is needed for coding the training data set, i.e., identifying critical media coverage of organizational actions. On the other hand, expertise is also needed in feature selection, i.e., in the identification and elimination of terms specific to the training data. In our case, these were terms (e.g., names) that are strongly associated with individual scandals that were addressed multiple times in our training dataset. Such context-specific terms help in the classification of the reference texts but decrease the quality of the classification of new texts (overfitting problem).

Using an example from organizational legitimacy research, we demonstrated that HACTA could be used to develop accurate classifiers. Our classifier, which identifies critical media coverage of organizations, is well capable of reproducing the manual coding of the training data set. Furthermore, the results of an external validation are positive, and a qualitative analysis indicates that our classifier produces only an insignificant number of false negatives and false positives.

However, the subcategories show limits of our approach: Our method can, obviously, only identify those types of criticized practices that were discussed in the reference texts. For example, our classifier is not able to identify environmental scandals due to the lack of an environmental category. Since the dictionary consists mainly of general terms (legitimacy and legality categories), our classifier would, nevertheless, classify articles that criticize environmental organizational behaviour correctly as “critical.”

Another possible point of criticism is the bag of words-assumption that underlies all three approaches presented as it ignores word order information and, thus, semantic relationships between words. One way to address this issue would be to add n-grams, i.e., a contiguous sequence of words (e.g., bigrams like “very good” or trigrams like “not very good”) to the analysis. However, studies have shown that this does not necessarily result in better classifiers (Kobayashi et al., 2018a). In general, bag-of-words approaches work well when it can be assumed that text categories differ strongly in the use of individual terms. For a large part of text classification tasks, this is the case, especially for topic classification tasks. This is why the bag-of-words approach has proven to work well here in particular (HaCohen-Kerner et al., 2020). However, this is not always the case: Especially in sentiment analysis, the use of bag-of-words approaches has been criticized in the recent past (e.g., Rudkowsky et al., 2018), since negations (e.g., "good" vs. "not good" or "not very good"), for example, are not correctly recognized.

For this reason, natural-language processing (NLP) techniques are increasingly being used in this area. Here, the analysis is shifted to the sentence level: for example, part-of-speech tagging is used to identify words as nouns, verbs, adjectives, etc., and by using dependency parsing, even the grammatical structure of a sentence can be determined. These approaches have been frequently used to improve accuracy in sentiment analysis (e.g., Chan & Chong, 2017). A rather new approach is the use of distributed word embeddings (like word2vec) which represent words in a continuous vector space. In this way, relationships between words in textual data are detected: words occurring in the same context are considered semantically similar. Word embeddings thus can improve accuracy by identifying synonymous terms that are not or only scarcely represented in the training data. So far, word embeddings are mainly used for sentiment analysis, but they can also be used for other classification tasks (Rudkowsky et al., 2018). However, the use of word embeddings and other complex NLP techniques seems appropriate only where less extensive bag-of-words methods cannot achieve sufficient accuracy.

References

Agarwal, S., & Sureka, A. (2016). *But i did not mean it!—intent classification of racist posts on tumblr.* Paper presented at the 2016 European Intelligence and Security Informatics Conference (EISIC).

Aggarwal, C. C. (2018). *Machine learning for text*. New York: Springer.

Ahmad, K., Han, J., Hutson, E., Kearney, C., & Liu, S. (2016). Media-expressed negative tone and firm-level stock returns. *Journal of Corporate Finance, 37*, 152–172.

Andal-Ancion, A., Cartwright, P. A., & Yip, G. S. (2003). The digital transformation of traditional business. *MIT Sloan Management Review, 44*(4), 34–41.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). *Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.* Paper presented at the Lrec.

Bae, Y., & Lee, H. (2012). Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers. *Journal of the American society for information science and technology, 63*(12), 2521–2535.

Bannier, C., Pauls, T., & Walter, A. (2019). Content analysis of business communication: introducing a German dictionary. *Journal of Business Economics, 89*(1), 79–123.

Bansal, P., & Gao, J. (2006). Building the future by looking to the past: Examining research published on organizations and environment. *Organization & environment, 19*(4), 458–478.

Bednar, M. K. (2012). Watchdog or lapdog? A behavioral view of the media as a corporate governance mechanism. *Academy of Management Journal, 55*(1), 131–150.

Bernard, H. R., Wutich, A., & Ryan, G. W. (2016). *Analyzing qualitative data: Systematic approaches*: SAGE publications.

Bonfiglioli, R., & Nanni, F. (2015). *From close to distant and back: how to read with the help of machines.* Paper presented at the 3rd International Conference on the History and Philosophy of Computing, Pisa, Italy.

Caserio, C., Panaro, D., & Trucco, S. (2019). Management discussion and analysis: a tone analysis on US financial listed companies. *Management Decision, 58*(3), 510–525.

Castelló, I., Etter, M., & Årup Nielsen, F. (2016). Strategies of legitimacy through social media: The networked strategy. *Journal of management studies, 53*(3), 402–432.

Chan, S. W., & Chong, M. W. (2017). Sentiment analysis in financial texts. *Decision Support Systems, 94*(2), 53–64.

Cogburn, D., & Hine, M. (2017). *Introduction to Text Mining in Big Data Analytics Minitrack.* Paper presented at the Proceedings of the 50th Hawaii International Conference on System Sciences, HICSS 2017.

Durand, R., & Vergne, J. P. (2015). Asset divestment as a response to media attacks in stigmatized industries. *Strategic Management Journal, 36*(8), 1205–1223.

Duriau, V. J., Reger, R. K., & Pfarrer, M. D. (2007). A content analysis of the content analysis literature in organization studies: Research themes, data sources, and methodological refinements. *Organizational research methods, 10*(1), 5–34.

Evans, M., McIntosh, W., Lin, J., & Cates, C. (2007). Recounting the courts? Applying automated content analysis to enhance empirical legal research. *Journal of Empirical Legal Studies, 4*(4), 1007–1039.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*: Cambridge university press.

Garrad, M. W. (2003). *Computer Aided Text Analysis in Personnel Selection.* PhD-thesis, School of Applied Psychology, Griffith University,

George, G., Osinga, E. C., Lavie, D., & Scott, B. A. (2016). Big data and data science methods for management research. *Academy of Management Journal, 59*(5), 1493–1507.

Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications, 40*(16), 6266–6282.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly, 93*(2), 332–359.

HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PloS one, 15*(5), 1–22.

Harish, B. S., Guru, D. S., & Manjunath, S. (2010). Representation and classification of text documents: A brief review. *IJCA, Special Issue on RTIPPR (2)*, 110–119.

Hausladen, C. I., Schubert, M. H., & Ash, E. (2020). Text classification of ideological direction in judicial opinions. *International Review of Law and Economics, 62*, 1–19.

Heitmann, M., Siebert, C., Hartmann, J., & Schamp, C. (2020). *More than a feeling: Benchmarks for sentiment analysis accuracy*. Ssrn, Working Paper. Retrieved 29.3.2021, from https://dx.doi.org/10.2139/ssrn.3489963.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley Publishing Company.

Holton, C. (2009). Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem. *Decision Support Systems, 46*(4), 853–864.

Hossfeld, H. (2013). Corporate dieting. Persuasive use of metaphors in downsizing. *management revue, 24*(1), 53–70.

Hossfeld, H. (2018). Legitimation and institutionalization of managerial practices. The role of organizational rhetoric. *Scandinavian Journal of Management, 34*(1), 9–21.

Huang, A. H., Zang, A. Y., & Zheng, R. (2014). Evidence on the information content of text in analyst reports. *The Accounting Review, 89*(6), 2151–2180.

Humphreys, A., & Wang, R. J.-H. (2018). Automated text analysis for consumer research. *Journal of Consumer Research, 44*(6), 1274–1306.

Ignatow, G., & Mihalcea, R. (2016). *Text mining: A guidebook for the social sciences*. Thousan Oaks, California: Sage Publications.

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS transactions on computers, 4*(8), 966–974.

Kabanoff, B. (1997). Introduction: computers can read as well as count: computer-aided text analysis in organizational research. *Journal of Organizational behavior, 18*, 507–511.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018a). Text classification for organizational researchers: A tutorial. *Organizational research methods, 21*(3), 766–799.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018b). Text mining in organizational research. *Organizational research methods, 21*(3), 733–765.

Kothari, S. P., Li, X., & Short, J. E. (2009). The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review, 84*(5), 1639–1670.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American political science review, 97*(2), 311–331.

Laver, M., & Garry, J. (2000). Estimating policy positions from political texts. *American Journal of Political Science*, 619–634.

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of broadcasting & electronic media, 57*(1), 34–52.

Loughran, T., & McDonald, B. (2015). The use of word lists in textual analysis. *Journal of Behavioral Finance, 16*(1), 1–11.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research, 54*(4), 1187–1230.

Martin, A. D., Quinn, K. M., Ruger, T. W., & Kim, P. T. (2004). Competing approaches to predicting supreme court decision making. *Perspectives on Politics, 2*(4), 761–767.

Park, J. r., Lu, C., & Marion, L. (2009). Cataloging professionals in the digital environment: A content analysis of job descriptions. *Journal of the American society for information science and technology, 60*(4), 844–857.

Pencle, N., & Mălăescu, I. (2016). What's in the words? Development and validation of a multidimensional dictionary for CSR and application using prospectuses. *Journal of Emerging Technologies in Accounting, 13*(2), 109–127.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Retrieved 29.3.2021, from https://mcrc.journalism.wisc.edu/files/2018/04/Manual_LIWC.pdf.

Pérez-Vera, S., Alfaro, R., & Allende-Cid, H. (2017). *Intent classification of social media texts with machine learning for customer service improvement.* Paper presented at the International Conference on Social Computing and Social Media.

Pfeffer, J., & Gerald, R. (1978). *The external control of organizations: A resource dependence perspective*: New York: Harper & Row.

Piccarozzi, M., Aquilani, B., & Gatti, C. (2018). Industry 4.0 in management studies: A systematic literature review. *Sustainability, 10*(10), 1–24.

Platanou, K., Mäkelä, K., Beletskiy, A., & Colicev, A. (2018). Using online data and network-based text analysis in HRM research. *Journal of Organizational Effectiveness: People and Performance*.

Pollach, I. (2012). Taming textual data: The contribution of corpus linguistics to computer-aided text analysis. *Organizational research methods, 15*(2), 263–287.

Prüfer, J., & Prüfer, P. (2018). Data science for institutional and organizational economics. In C. a. M. M. S. Ménard (Ed.), *A research agenda for new institutional economics* (pp. 248–259). Cheltenham: Edward Elgar Publishing.

Riffe, D., Lacy, S., Fico, F., & Watson, B. (2019). *Analyzing media messages: Using quantitative content analysis in research*: Routledge.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures, 12*(2–3), 140–157.

Scherer, A. G., & Palazzo, G. (2011). The new political role of business in a globalized world: A review of a new perspective on CSR and its implications for the firm, governance, and democracy. *Journal of management studies, 48*(4), 899–931.

Sheng, J., Amankwah-Amoah, J., & Wang, X. (2017). A multidisciplinary perspective of big data in management research. *International Journal of Production Economics, 191*, 97–112.

Strycharz, J., Strauss, N., & Trilling, D. (2018). The role of media coverage in explaining stock market fluctuations: Insights for strategic financial communication. *International Journal of Strategic Communication, 12*(1), 67–85.

Stulpe, A., & Lemke, M. (2016). Blended reading. In M. Lemke & G. Wiedemann (Eds.), *Text Mining in den Sozialwissenschaften* (pp. 17–61). Wiesbaden: Springer.

Thomson Reuters (2017). *Thomson Reuters ESG Scores*. Retrieved 29.3.2021, from https://www.esade.edu/itemsweb/biblioteca/bbdd/inbbdd/archivos/Thomson_Reuters_ESG_Scores.pdf.

Tonidandel, S., King, E. B., & Cortina, J. M. (2018). Big data methods: Leveraging modern data analytic techniques to build organizational science. *Organizational research methods, 21*(3), 525–547.

Vaara, E., & Tienari, J. (2002). Justification, legitimization and naturalization of mergers and acquisitions: A critical discourse analysis of media texts. *Organization, 9*(2), 275–304.

van Zoonen, W., & Toni, G. (2016). Social media research: The application of supervised machine learning in organizational communication research. *Computers in Human Behavior, 63*, 132–141.

van Zoonen, W., Verhoeven, J. W., & Vliegenthart, R. (2016). How employees use Twitter to talk about work: A typology of work-related tweets. *Computers in Human Behavior, 55*, 329–339.

Watanabe, K. (2016). *Computer-aided dictionary making: An efficient dictionary construction technique for content analysis.* Paper presented at the Proc. Int. Conf. on the Advances in Computational Analysis of Political Text.

Watanabe, K. (2020). Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages. *Communication Methods and Measures*, 1–22. Retrieved 29.3.2021, from https://doi.org/10.1080/19312458.2020.1832976.

Wenzel, R., & Van Quaquebeke, N. (2018). The double-edged sword of big data in organizational and management research: A review of opportunities and risks. *Organizational research methods, 21*(3), 548–591.

Xie, P., & Xing, E. (2018). *A neural architecture for automated icd coding.* Paper presented at the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Yehia, A. M., Ibrahim, L. F., & Abulkhair, M. F. (2016). Text mining and knowledge discovery from big data: challenges and promise. *International Journal of Computer Science Issues (IJCSI), 13*(3), 54.

Appendix: Main R Packages and Functions Used by the Authors

Task	R-package	Use of specific functions
Extracting text data into a single data set	base	■ misc. (to create a matrix with documents as rows)
Cleaning text data	base, tm	■ “gsub” (to remove numbers, special characters, and white space) ■ “tolower“ (to convert to lowercase) ■ “tm_map“ (to remove stop words)
Creating a DTM	tm	■ “DocumentTermMatrix“
Reduction of sparsity	tm	■ “RemoveSparseTerms”
Calculating selective terms	base	■ misc. (to calculate absolute and expected values)
Creating a dictionary	quanteda	■ “dictionary”
Applying a dictionary	quanteda	■ “dfm”