Title: Business analytics for strategic management: Identifying and assessing corporate challenges via topic modeling

Author: Nicolas Pröllochs Stefan Feuerriegel

PII: S0378-7206(17)30925-4
DOI: https://doi.org/doi:10.1016/j.im.2018.05.003
Reference: INFMAN 3070

To appear in: *INFMAN*

Received date: 20-10-2017
Revised date: 2-5-2018
Accepted date: 4-5-2018

Please cite this article as: Nicolas Pr*ddoto*llochs, Stefan Feuerriegel, Business analytics for strategic management: Identifying and assessing corporate challenges via topic modeling, *<![CDATA[Information & Management]]>* (2018), https://doi.org/10.1016/j.im.2018.05.003

Business analytics for strategic management: Identifying and assessing corporate challenges via topic modeling

Nicolas Pröllochs[a,*], Stefan Feuerriegel[b]

[a] *Chair for Information Systems Research, University of Freiburg, Platz der Alten Synagoge, 79098 Freiburg, Germany*
[b] *ETH Zurich, Weinbergstr. 56/58, 8092 Zurich, Switzerland*

Abstract

Strategic management requires an assessment of a firm's internal and external environments. Our work extends the body of management tools (e.g., SWOT analysis or growth-share matrix) by proposing an automated text mining framework. Here we draw on narrative materials from firms (e.g., financial disclosures) and perform topic modeling so as to identify the key issues faced by an organization. We then quantify the use of language along two dimensions: risk and optimism. This reveals a firm's strengths and weaknesses by identifying business units, activities, and processes subject to risk, while also comparing it with competitors or the market.

Keywords: Business analytics; Text mining; Firm performance; Topic modeling; Latent Dirichlet Allocation; Strategic management

Highlights

- Strategic management relies upon internal and external performance assessments of firms
- We propose a model based on studying the language in narrative materials

*Corresponding author. Tel: +49 761 203 2396; Fax: +49 761 203 2416.
Email addresses: nicolas.proellochs@is.uni-freiburg.de (Nicolas Pröllochs), sfeuerriegel@ethz.ch (Stefan Feuerriegel)

- It combines topic modeling with a topic-specific analysis of risk/opportunity terms
- Advantages are automation, a high granularity and the lack of subjective judgments
- It monitors internal risks/strengths and, externally, compares them to the market

Acknowledgments

The valuable contributions of Ryan Grabowski, Anton Mosin, Dirk Neumann and Georg Wolff are gratefully acknowledged.

Business analytics for strategic management: identifying and assessing corporate challenges via topic modeling

Abstract

Strategic management requires an assessment of a firm's internal and external environments. Our work extends the body of management tools (e.g., SWOT analysis or growth-share matrix) by proposing an automated text mining framework. Here we draw on narrative materials from firms (e.g., financial disclosures) and perform topic modeling so as to identify the key issues faced by an organization. We then quantify the use of language along two dimensions: risk and optimism. This reveals a firm's strengths and weaknesses by identifying business units, activities, and processes subject to risk, while also comparing it with competitors or the market.

Keywords: Business analytics; Text mining; Firm performance; Topic modeling; Latent Dirichlet allocation; Strategic management

1. Introduction

A firm's strategy undergoes regular updates, whereby the understanding of firm internals or its environment changes. In this sense, strategic management consists of analyses, decisions, and actions undertaken by organizations to create or sustain a competitive advantage [1]. Strategies are assumed to be responsive and not static, as they include a feedback loop to monitor progress and adapt subsequent planning to it [2]. For this reason, it is key to have accurate knowledge of what occurs in both the internal environment and the external environment.

To facilitate the task of strategic analysis, academics and practitioners have devised a variety of management tools [3, 4], each with a different objective [1]. We list a few illustrative examples. In terms of an external perspective, the PESTLE framework performs a macroanalysis of political, economic, social, technological, legal, and environmental factors. Analogously, the growth-share matrix facilitates resource allocation by visualizing products or business units in terms of market shares and growth rates. As an internal analysis, the balanced scorecard provides a semistructured report tracking the progress with which activities are executed. Both the internal view and the external view are taken into account by SWOT analysis to identify current and future performance. Accordingly, it derives key factors for accomplishing a strategic goal, grouped by strengths, weaknesses, opportunities, and threats. Since in this work we are interested in a holistic assessment involving both internal and external factors, we later delve into the ideas of SWOT analysis and draw on dimensions related to both risk and performance outlook. This work specifically draws on the metrics of

SWOT analysis as they—despite their age—still enjoy widespread application in business planning, in management practice, and as a core vehicle for management consulting firms, thereby yielding direct value to practitioners and businesses [5, 4].

The aforementioned frameworks were developed at a time when computers were not as ubiquitous as today, and hence they entail several inherent drawbacks: they reflect the outcome of a single event, as they must be updated manually without automation [1]; they tend to assess firms using discrete scales but cannot live up to the granular precision of continuous scores; and they cannot prescribe dimensions on the basis of which firms are ranked, thus running the risk of not identifying the relevant metrics [6].

To address the aforementioned limitations, we translate a risk-optimism analysis similar to the SWOT matrix into a computerized procedure. For this purpose, our approach draws on firm-related narratives such as regulatory disclosures, financial filings, press releases, or news coverage. Depending on which text source is chosen, our framework either uncovers the (perceived) internal workings of firms or reflects an external view. It then follows a two-stage process. It first uses topic modeling to identify clusters within the narrative materials. These reflect the management-relevant undertakings within firms and often pertain to current challenges or performance issues. In a second step, we assign numerical scores to the linguistic tone depending on predefined dimensions. In our case, we use the theoretical foundation from [7] and thus look for terms that signal risks and optimism, as such terms are most likely to engage the attention of the top management and will most probably trigger subsequent actions. This enables practitioners to

map performance indicators onto a risk-optimism matrix that presents the outcome in a highly visual fashion.

Figure 1: Illustrative visualization of our strategic framework based on the linguistic style. Three examples are plotted together with their use of natural language in describing legal subjects.

Our risk-optimism framework provides monitoring capacity as a management tool. Figure 1 presents an illustrative example in which three firms are analyzed with respect to the legal content of their press coverage and examples of characteristic wordings are identified. These are then translated into risk-optimism scores. Our approach allows one to quantify the risk-related tone and compare individual firms with competitors, thereby guiding management toward a more risk-focused strategy. At the same time, it also helps those responsible for financial management before they make investments. As another example, we can monitor how competing firms perceive their own expansion into new markets. Whether such initiatives actually yield financial success is often unclear for a long time. Our tool helps to quantify the firm's view as the tool numerically ranks the different markets. Similarly, companies can monitor expansions by competitors so as to gain an understanding of how successful different markets have turned out to be, while financial figures are often confidential at such a granular level.

While management tools and business analytics have been developed separately hitherto, it is of interest whether the benefits of the two procedures can be combined . Hence, this yields our research question as follows: can we translate traditional management tools into computerized procedures? What firm-related issues can be extracted through content analysis? For this pur-

pose, we present a data-driven approach to the classical frameworks used in strategic management. Our study thereby demonstrates how firms can leverage advanced analytics as a monitoring tool. Here we process unstructured data, which corresponds to the dimension of "variety" in IBM's five Vs of big data and, therefore, fulfills the corresponding definition of big data analytics [8].

We demonstrate empirical findings for our proposed method by using an empirical setting from the energy sector. We specifically decided on this market segment as climate change, corresponding policies, and liberalization efforts have forced energy companies to simultaneously adapt their strategy in various directions [9]. This includes, for instance, direct changes to their energy consumption, redefining sources of energy generation, and reducing emissions. Accordingly, existing industries (such as those involving fossil fuels or nuclear power) are threatened in many countries, while new players (e.g., decentralized microgrids) are entering the market. These challenges have been identified by information systems research, which has resulted in the field known as *green information systems* [10]. At the same time, sustainability itself has become a strategic goal for firms [11, 12]. Hence, our empirical study investigates whether our strategic framework can identify relevant risks and challenges in this sector. For this purpose, we use a dataset of regulated financial disclosures from the US market to identify the relevant themes in the energy sector. We find that most documents can be assigned to a relatively small set of topics. However, a thorough analysis based on the aforementioned risk-optimism framework reveals that the tone differs substantially across the identified topics. For instance, we find that financial

disclosures related to earnings results show a highly optimistic tone, whereas filings related to production outlook show a relatively high risk level add. This also confirms and extends finding from other studies [e.g. 13, 14].

Our management framework pertains to both the internal view and the external view of companies. On the one hand, it surveys different areas, initiatives, and developments within a firm. Thereby, one can identify those companies whose risk score exceeds the desired threshold as a direct managerial implication. On the other hand, managers can also compare their firm with direct competitors or even with the market as a whole. One can compare a firm's performance with that of competitors in core areas where financial figures are difficult to come by, as, for instance, showcased in the preceding examples. This thus entails various implications for practitioners (e.g., serving as an early warning mechanism).

Our work contributes to the field of strategic analysis by proposing a management tool with the following caveats. First, it requires no manual evaluations, but rather relies on computational routines. It thereby increases the speed with which such analyses are conducted. Second, our concept allows granular insight as it can provide recommendations at the level of business units, activities, and processes. Third, strategic management often conducts analysis along predetermined dimensions, while our method identifies common themes from the narratives themselves, thereby ensuring a holistic analysis.

We proceed as follows. Section 2 provides a background on topic modeling in the field of business analytics. Accordingly, we develop our computational routines in Section 3. We then use an empirical setup from the energy sector

(Section 4), on the basis of which Section 5 demonstrates our approach and reports empirical findings. Section 6 discusses implications for practitioners and managers, thereby detailing how our framework supports strategic management practice. Section 7 concludes the article.

2. Background

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 thus point towards different research streams concerning language-based business analytics. These are predominantly 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. These often consist of user-generated content, such as product reviews or social media, where sentiment analysis facilitates insights into the opinion of customers toward products. As illustrative examples, the opinion of individuals toward guns can be derived from the language in social media [15], while call center emails predict customer churn [16]. Alternatively, one can rely on perception as conveyed by professional reports, such as news or industry reports, or public materials from the firms themselves (e.g., regulatory filings, CEO speeches). These can help in improving the performance of investments [17, 18, 19]. Moreover, internal use of language (e.g., from emails) reflects the structure and processes of organizations. For instance, analysis of the relevant communication style

enables firms to identify fraud risk from disgruntled employees [20].

The applications of text mining also show considerable variations in terms of the underlying methods. On the one hand, various use cases require supervised learning with a priori labels. Examples include automatic assignment of IT tickets to the correct service unit [21], forecasts of news-based stock price changes [22], and predicting users' affect [23]. Alternatives rely on unsupervised methods, such as clustering or topic modeling, which are able to shed light on the patterns within business data. Illustrative demonstrations from businesses include measuring business proximity [24], predicting interest among tourists [25], and forming IT support groups on the basis of the content of helpdesk tickets [21].

3. Research framework

This section first details the intuition behind our computational procedures, followed by the necessary preprocessing of natural language and the topic-dependent language analysis.

3.1. Theoretical foundation

This article follows a practice-based view on strategy [26], whereby the goal is to help managers by focusing on actual techniques for developing strategies. This differs from another prominent, yet different, approach in strategic management: a resource-based view. The latter focuses on activities that cannot be adopted by other firms, whereas the practice-based view applies to our setting, as strategy is considered to be "imitable activities, or practices amenable to transfer across firms" [26].

Extensive research has been conducted on the outcomes of effective strategies, including the identification of environmental enablers of, and obstacles to, strategy. Related evidence from a large meta-analysis suggests that companies can outperform others by adopting a formal planning process [27]. However, this necessitates a fit between strategy and operating environment. Current work also focus on the dynamics of strategy-making, thereby incorporating temporal aspects [28].

Situational analysis refers to approaches that acquire information concerning two dimensions; namely, the firm's internal capabilities and the external environment consisting of opportunities and threats [7]. The eventual goal is to evaluate the internal view against the external view to identify actions that can change the firm's current condition. This is also known to facilitate new insights into the competition and reveal potential levers for increasing performance [29], as incorporated in the SWOT analysis as a management tool [4].

At the same time, research on organizational learning indicates a relationship between augmenting the strength of a firm and its performance in the market [30]. Organizational learning benefits especially from a fit between strategy and implementation [31]. In this context, organizational structure, and leadership in particular, plays a critical role; for instance, leadership alignment affects strategy implementation [32], while Vaara and Lamberg [28] discussed processes in the organizational structure necessary for the successful implementation of strategic changes. We refer to [33] for a discussion concentrating on behavioral aspects, as heuristic rules in decision-making introduce a form of systematic error.

3.2. Framework choice

The aforementioned research efforts in strategic management have led to a variety of management tools. We refer to [4] for a comprehensive overview. According to Alba and Hutchinson [7], the dimensions of risks and opportunities represent decisive instruments for strategic planning, and our work thus focuses on their operationalization in SWOT analysis. This choice is substantiated by a variety of practical arguments, as discussed in the following.

SWOT analysis has found widespread adoption in business planning, in management practice, and as a core vehicle for management consulting firms. Recent research has confirmed that "its use is routinized in many organizations" [4]. Examples include government bodies, such as the Queensland government[1] or the European Commission,[2] as well as public organizations, including UNICEF.[3] This matches common practice in top-tier strategy consulting firms where the SWOT analysis often underlies their work.

In light of the preceding arguments, we decided to translate SWOT analysis into computerized procedures as this promises the highest gain for businesses and direct impact in terms of management practice.

[1] https://www.business.qld.gov.au/starting-business/planning/market-customer-research/swot-analysis/uses

[2] http://forlearn.jrc.ec.europa.eu/guide/4_methodology/meth_swot-analysis.htm

[3] https://www.unicef.org/knowledge-exchange/files/SWOT_and_PESTEL_production.pdf

3.3. Method overview

We now develop our computational routines for tracking firm performance across different business units, processes, and activities. Figure 2 provides an overview, combining both topic extraction and a topic-specific language analysis. The idea of performing a language analysis has already been adopted in strategic management, where one has to manually search for trigger words that signal cost-conscious behavior of rivals [1].

Figure 2: Our proposed strategic framework, which first extracts relevant topics and then analyzes their language.

The key feature of our approach is that it incorporates narrative materials. Thereby, our framework differs from common alternatives in strategic management that use human surveys, subjective expertise, or decision-making heuristics. However, several unique features of natural language yield particular benefits: First, we overcome potential challenges that might arise when managers try to strategically ignore unfavorable issues. In contrast, our framework provides a holistic analysis of the strategic position. Second, our techniques provide a quantifiable metric that can easily be compared across organizations. Third, our technique can even detect subtle opinions such as the firm's subjective perception of policy risk.

We further note that linguistic materials can originate from a variety of sources, each with its own benefits. The use of regulatory filings puts emphasis on the internal reporting of firms and quantifies the management perspective of a firm's performance. Media articles rather shift the focus to the perceived performance of firms. This introduces the ability to register issues that might not have appeared in regulatory filings.

Topic modeling renders it possible for our framework to identify the different themes in the corpus [34]. As a result, we uncover the different issues that either are being disclosed by a firm or have been discussed by the media. We specifically advocate the use of topic modeling, since it provides an automated procedure for breaking down the corpus into themes. Moreover, this technique offers high flexibility, as it allows one to vary the number of topics, thus enabling control over the granularity of the analysis. As such, it can either establish general themes, such as industry units, or detail individual activities or processes. In contrast to naïve approaches in strategic management, our method does not prescribe specific corporate challenges or dimensions along which a firm is analyzed, but, instead, infers the relevant items directly from the data themselves. The conventional method for topic modeling is latent Dirichlet allocation (LDA) [35, 36]. It enjoys widespread use in research related to management (e.g., [25, 21, 24]), and in particular, is successful in extracting topics from financial materials [37].

Subsequent to topic modeling, we rate the individual documents in each topic according to two strategy-related dimensions: risk and optimism. Here we follow the intuition of the SWOT matrix and academic evidence [7]. The risk-optimism combination is further assumed to drive investment decisions in financial markets to a large extent. Hence, both dimensions strongly appeal to management, since a typical strategy aims at moving items along them: managers find a competitive advantage in well-performing items, while risk factors should be converted into nonrisk factors. The two dimensions specifically facilitate our goal of monitoring.

The actual rating of risk and optimism is performed by our analyzing the

linguistic styles within the documents. Extracting optimism and pessimism from disclosures or news has been a prominent theme in finance, and for this reason, researchers have devised accurate collections of linguistic expressions that imply an optimistic or pessimistic outlook [38, 13]. Similarly, corporations are known to disclose risk-related information in narrative materials, since this presents a means of avoiding litigation risks such as might arise in the process of initial public offerings [39]. Hence, this justifies our idea of quantifying optimism and risk by analyzing the use of natural language.

The conventional assumption [40, 41] is that these approaches can reliably infer the (subjective) information and encode it into a numerical rating with an accuracy sufficient for the given purpose. This assumption has been confirmed consistently, including in settings from accounting [13]. One commonly uses rule-based approaches for counting the occurrences of word labels a priori, according to the previous dimensions [40, 42]. Rule-based approaches are frequently implemented when financial materials are being processed because of their objectivity and reliability [43, 44, 13]. We then later count the frequency of terms indicating optimism or risk to obtain a single score. Since we are not aware of a collection of risk words, we describe its construction in Appendix A.

Our procedure resembles common approaches in natural language processing [45, 41], and we briefly mention the relationships in the following. For instance, sentiment analysis quantifies the use of positive and negative language in texts [41], whereas we specifically propose the assessment along other dimensions. Affective computing extends the previous dimensions and sets out to measure the subjective emotions of natural language [46]. An-

other approach stems from aspect-based sentiment analysis, which focuses on a specific entity and senses the polarity of narratives toward it [40].

3.4. Preprocessing

In a first step, we preprocess the narrative materials following common operations in text mining [45, 41, 42]. This initial data preparation includes standard routines from the domain of natural language processing to transform the running text into a structured format that allows further calculations. Specifically, we use a list of cutoff patterns to extract only the textual components from the documents (i.e., we remove contact addresses and HTML/XML formatting), omit stop words without a deeper meaning, and use Porter's stemming algorithm to truncate inflected words to their stem [45]. The resulting word frequencies are highly skewed toward zero, since most words appear only in a subset of all documents. For this reason, we adhere to the suggestions in [47] and omit words that occur in less than 1 % of all documents. Then, t refers to an arbitrary word and tf_t to its frequency in the corpus $\mathcal{D}$.

3.5. Topic modeling

Topic modeling provides a statistical means for detecting latent themes within a collection of documents [34]. Accordingly, this technique draws on word frequencies and then groups documents into clusters with similar content. LDA is a common probabilistic model [36]. As a main advantage, this method is based on highly efficient probability inference algorithms, and has been found to yield highly interpretable topics in an unsupervised fashion [36]. In its mathematical representation, each topic is a distribution over

words in the vocabulary, and every document is modeled as a distribution over topics. Accordingly, every document is assumed to have been generated by the following process [35, 36]:

1. *Document-topic relationship.* For every document d in corpus $\mathcal{D}$, draw a random variable $\theta_d \in \mathbb{R}^K$ from the Dirichlet distribution given by $\theta_d \sim \text{Dir}(\alpha)$. Here θ_d specifies the relative proportion with which the K topics appear in a given document.

2. *Word frequency.* For each topic k, draw a random variable $\beta_k \sim \text{Dir}(\eta)$, which specifies the distribution of terms in that specific topic.

3. *Topic-word relationship.* For every word t in document d, draw a topic $z_t \sim \text{Mult}(\theta_d)$ from a multinomial distribution with θ_d prior and a scaled word frequency $tf_t \sim \text{Mult}(\beta_{z_t})$ from the multinomial distribution.

Accordingly, the purpose of LDA is to estimate the posterior distribution of topics β and topic proportions θ. Then, $P(tf_t \mid z_k, \beta)$ denotes the probability that a word t occurs in a chosen topic z_k. Consistent with previous research (e.g., [37]), we assign each document to the topic with the highest probability. As a main benefit, this strategy simplifies the interpretation of our later language analysis, and also lends itself to the unique nature of 8-K filings, which typically inform investors about a single topic (such as earnings results or management changes). A sensitivity analysis yielded empirical results in favor of this strategy, since most documents have one topic with a probability of more than 50 %.[4]

[4]We also experimented with an alternative configuration that allows each document to have multiple topics on the basis of the posterior probability. This approach yields

To this end, the joint likelihood for estimating the model is

$$P(\theta, \beta, tf, z) = \prod_{d=1}^{|\mathcal{D}|} P(\theta_d \,|\, \alpha) \prod_{k=1}^{K} P(\beta_k \,|\, \eta) \prod_{t} P\big(z_t^d \,\big|\, \theta_d\big)\, P\big(tf_t^d \,\big|\, z_t^d\big). \tag{1}$$

Directly estimating the model is computationally intractable, and as a remedy, one relies on approximate inference techniques such as variational expectation maximization [48]. The Dirichlet priors α and η control the document-topic and topic-word distributions, respectively. We initialize all LDA parameters by using the default values used in the original article by Blei et al. [35].

Finally, the LDA requires one to assign a unique identifier (i.e., topic name) to each of the extracted topics. To interpret a topic, one typically examines a ranked list of the 3 to 30 most probable terms in that topic. As a drawback, frequent and indecisive terms in the corpus commonly appear in such lists, and hence render it difficult to differentiate the meanings of the topics. Consequently, recent research found that ranking terms based on this probability hampers interpretation (e.g., [49]). To mitigate this issue, we use the term-topic relationship scheme from Sievert and Shirley [50], which facilitates topic interpretation by measuring the *relevance* of a term to a topic. From a mathematical perspective, *relevance* is a weighted average of the logarithms of a term's probability and the ratio of a term's probability within a topic to its marginal probability across the corpus. As a main benefit, this method results in more coherent and interpretable topics [50].

qualitatively equal results that are available on request.

3.6. Language analysis

In a second stage, we measure the use of linguistic terms pertaining to the different performance dimensions; namely, risk and optimism. For this purpose, we follow a rule-based approach that counts the frequency of corresponding terms as defined in predetermined word lists (see Appendix A). Thereby, we obtain an approach that is computationally efficient, requires little supervision in terms of labeled data, and achieves a high explanatory power.

We measure the orientation of natural language with regard to risk and optimism on the basis of the following rules. Let R_d denote the resulting risk and O_d the optimism score of document d. We then compute a simple ratio of the number of labeled words to the total number of words; that is,

$$R_d = \frac{\text{number of risk words}}{\text{total number of words}} \text{ and } O_d = \frac{\text{number of optimism words} - \text{number of pessimism words}}{\text{total number of words}}. \tag{2}$$

Such simple ratios are commonly used in the mining of textual materials from accounting because of their reliability and robustness [13]. Nevertheless, practitioners might prefer alternatives that further extend linguistic terms t with a metric w_t discriminating the relative importance (e.g., "risk" may be a stronger indicator of risk than "may"). The benefit of our approach is that it can be easily replaced by such a scoring mechanism; that is, changing R_d to $\sum_t w_t \, tf_t$/total number of words, and changing O_d analogously [51]. For reasons of reproducibility and ease of use, we demonstrate the outcomes for a simple categorization without weighting in the following. In addition, for many documents, only a few words are labeled as risk-related, optimistic, or

pessimistic expressions, thereby resulting in values scattered closely around zero. For this reason, we follow prevailing conventions [38] and finally standardize the corresponding scores to zero mean and a standard deviation of 1 to allow better comparability.

Alternatively, one could draw on supervised machine learning [41]; however, this would require labeled data, which again would explicitly incorporate human judgments with regard to the labels. Moreover, it is not clear whether the classifiers would make accurate predictions based on the language and expressions from an unseen topic after having been trained only on other, domain-specific language. For instance, even deep neural networks or their combination with transfer learning hardly surpasses the performance of rule-based classifiers [52, 19]. We thus leave this to future research.

4. Empirical setup

4.1. Study setting

Our empirical setup draws specifically from the energy sector, since it is undergoing a major transition in many countries. This entails extensive changes, including the gradual reduction of greenhouse gases emissions, and new corporations entering the field, thereby impacting operations and strategy [9]. For instance, investment decisions are governed by considerations of risk and return, and yet they are actively influenced by public policies and thus pose a perceived risk for investors [53].

We now describe the motivation for the choice behind our corpus. We chose financial filings on purpose, since regulations require firms to present a relatively unbiased picture of their current performance and accompanying

risks. Disclosures are further suited to identifying the current challenges facing organizations.

We retrieved Form 8-K filings from the US Securities and Exchange Commission. Firms are required to fill out Form 8-K to inform stakeholders about all recent developments and corporate events deemed relevant. This includes the financial performance, sales of securities, takeovers, and management changes, but also information specific to the energy sector, such as legal risks related to policy changes. Naturally, US regulations also oblige firms to announce unspecified events, such as the consequences of the Deepwater Horizon incident on April 20, 2010. Notably, this type of filing differs from that of Form 10-K and Form 10-Q, which disclose annual and quarterly earnings reports, respectively. In addition, Form 8-K must be filed in a timely fashion (currently four business days); for this reason, our method enables the taking of strategic action by the management without a long time delay. Finally, the filings also specify the sector in which the firm operates, which allows our framework to directly identify and filter firms relevant to our empirical setup.

4.2. Summary statistics

Our dataset spans the time frame from 2004 to 2017 and covers all stocks that were publicly traded on the New York Stock Exchange (NYSE).[5] This amounts to a total of 250,556 filings, which then undergo several filtering steps. First, to obtain information about the stock market reaction of investors, we remove filings for which we are not able to match the Securi-

[5]For reasons of reproducibility, the dataset is publicly available via `https://www.github.com/anonymized`.

ties and Exchange Commission Central Index Key numbers to NYSE stock symbols. Second, consistent with Loughran and McDonald [47], we exclude filings that contain fewer than 200 words. These filtering steps result in a final corpus of 249,481 filings, of which 29,874 (i.e., 12.0 %) belong to the energy sector.

We now investigate the frequency and length of the filings in the following. Our corpus covers a total of 177 different companies that operate in a wide range of energy subsectors, such as coal mining (e.g., Alpha Natural Resources Inc.) or offshore drilling (e.g., Transocean Ltd.). In addition to mid-size companies, this also includes filings from global players such as the world's largest oil field services company, Schlumberger Ltd. The median number of filings per firm is 132 (with a standard deviation of 91.09). The total range is from a minimum of 1 to a maximum of 477 for a single firm. The mean length of an individual filing totals 3,886.96 words. Figure 3 compares the frequency of filings across each year, revealing a slight trend over time.

Figure 3: Number of 8-K filings across the years 2004–2017. The white bars show the absolute number of filings in our study period, whereas the shaded bars correspond to filings of companies from the energy sector.

5. Empirical findings

In this section we analyze the previous empirical setting. We first perform topic modeling to identify the themes in the narrative materials from firms in the energy sector. For each topic, we then analyze the use of terms conveying information regarding risks or optimism. This forms the basis for further investigating internal and external performance.

5.1. Identification of topics

To perform the LDA, one has to choose ex ante the number of topics that one wants to identify. This differs from other machine learning algorithms whereby one optimizes, for example, the number of clusters by cross-validation or heuristics. Hence, we performed an experiment in which we presented the top 10 most relevant words of topic models with a different number of topics ($k = 5, 10, 20, 30, 40$) to five students from our business department. We then asked them to assign names to each topic and to decide which choice of k leads to individual topics being highly coherent and collectively exhaustive but mutually exclusive to other topics. The students consistently agreed on $k = 20$ topics, which is also concordant with related research [35, 54, 55]. As a feature of our management tool, one can vary the number of topics and thereby control the granularity of the analysis. We thus perform further empirical evaluations with other configurations as part of our sensitivity analysis, obtaining affirmative results that support our insights (see Section 6 for further details).

Subsequently, we use the feedback from our experiments and follow a two-stage approach to assign a unique name to each of the extracted topics. First, we infer the individual topic names from the most relevant terms occurring in each given topic. For example, stemmed words (such as *director*, *appoint*, *vote*, and *elect*) suggest a topic related to changes in management or corporate governance. Similarly, word stems such as *quarter*, *incom*, and *earn* indicate disclosures that are related to earnings results. Second, we thoroughly verify the topic names by manually assessing example filings from our dataset. We then asked our students once again to assign the inferred topic

names to the top 10 most relevant words. This process resulted in a large degree of agreement between the five students as measured by a relatively high interrater reliability in terms of Fleiss's kappa of 0.544. The individual topics are presented in Table 1, while a complete list of frequent terms for each topic is provided in Appendix B.

5.2. Descriptive statistics of topics

We now study the frequency and reception of the individual topics in more depth. As previously mentioned, we assign each filing in our sample to the topic with the highest posterior probability. Table 1 shows the frequency of each topic. Most documents are assigned to six topics; namely, *loan arrangement*, *trust indenture*, *earnings results*, *production outlook*, *mergers and acquisitions* and *dividend payment*. These six topics account for two-thirds of all filings, whereas the rest are distributed across the remaining ones. The high share of, for example, *earnings results* as a frequent topic in financial disclosures is also consistent with findings from the existing literature (e.g., [56]).

Table 1 also rates the individual topics according to the aforementioned strategic dimensions; namely, risk and optimism. For example, the fourth column denotes the mean risk score for 8-K filings that correspond to a particular topic. Evidently, both the median and the standard deviation differ substantially across the identified topics. For example, *production outlook* features a relatively high mean risk level (0.73), whereas *earnings results* conveys a lower risk (-0.80) on average. We observe a similar pattern regarding the optimism metric. Here, for instance, *management change* conveys high optimism (0.59), whereas *purchase agreement* features a relatively

No.	Topic name	No. of filings	Risk			Optimism		
			Mean	Median	SD	Mean	Median	SD
1	Loan arrangement	5086	−0.31	−0.35	1.00	0.32	0.36	0.89
2	Trust indenture	2542	0.70	0.71	0.69	−0.09	−0.07	0.60
3	Legal issues	110	−1.03	−0.93	0.43	−0.37	−0.52	0.68
4	Earnings results	1515	−0.80	−0.86	0.84	0.60	0.71	0.75
5	Income statements	150	−0.81	−0.83	0.83	−0.33	−0.53	0.96
6	Security agreement	73	0.09	−0.05	0.96	−1.12	−0.91	1.03
7	Employment agreement	452	0.40	0.40	0.80	−0.37	−0.27	1.18
8	Purchase agreement	831	0.50	0.62	0.82	−1.38	−1.42	1.49
9	Tax report	350	−0.24	−0.29	0.59	−1.84	−1.93	0.84
10	Stock option award	342	−0.81	−0.98	1.02	0.17	0.08	0.80
11	Resource development	492	−0.53	−0.64	0.69	−1.36	−1.23	0.97
12	Management change	979	0.67	0.65	0.89	0.59	0.61	0.88
13	Amendment of shareholder rights	728	0.21	0.30	0.65	−0.06	−0.15	0.61
14	Production outlook	2970	0.73	0.63	0.80	0.09	0.13	0.90
15	Infrastructure and logistics	541	−0.59	−0.63	0.79	0.07	0.16	0.84
16	Partnership arrangement	728	−0.63	−0.84	0.71	−0.19	−0.26	0.57
17	Mergers and acquisitions	1401	−0.77	−0.82	0.62	−0.12	−0.20	0.95
18	Public relations	758	−0.66	−0.70	0.38	−0.95	−1.01	0.52
19	Dividend payment	1556	0.44	0.38	0.86	0.35	0.37	0.71
20	Drilling contracts	1094	0.33	0.36	0.81	0.01	0.01	0.69

SD, standard deviation.

Table 1: Summary statistics across identified topics. Each filing is assigned to the topic with the highest posterior probability.

low optimism level (−1.38) on average. We see that the risk and optimism scores also feature different degrees of volatility with respect to the individual topics. For example, *purchase agreement* shows a standard deviation of 1.49 regarding the optimism metric, while *amendment of shareholder rights*, for example, features a lower value of 0.61. We discuss interpretations in Section 5.6.

5.3. Strategic analysis of the internal environment

Our approach enables practitioners to map performance indicators onto a risk-optimism matrix that presents the outcome in a highly visual fashion.

In the following, we present two examples that analyze all filings published by two distinct companies within our study period between 2004 and 2017.

Figure 4 presents an illustrative example for the world's largest publicly traded international oil and gas company, ExxonMobil.[6] This company holds an industry-leading inventory of resources and is one of the world's largest integrated refiners, marketers of petroleum products, and chemical manufacturers. Our framework immediately signals differences regarding the communication of different news topics. This allows companies to identify strengths and weaknesses from an internal perspective. For instance, filings related to *mergers and acquisitions* feature a pessimistic tone with a medium risk level. In contrast, ExxonMobil publishes press releases related to *dividend payment* and *earnings results* with a highly optimistic tone and a relatively low risk level. This is plausible, since ExxonMobil more than doubled its market capitalization during our study period. In addition, it was the second most profitable company in the Fortune 500 in 2014.

Figure 4: Risk-optimism matrix for news filings from ExxonMobil.

Next, we contrast the preceding analysis for ExxonMobil with that for another company from the energy sector. For this purpose, Figure 5 presents an additional risk-optimism matrix that analyzes the same topics for one of the world's largest offshore drilling contractors, Transocean Ltd. This company operates in more than 20 countries and provides offshore contract drilling services for oil and gas wells worldwide. As reflected in Figure 5,

[6]For space reasons, we limit the visualization to the six most common topics, and a complete risk-optimism matrix for all topics is available on request.

we immediately observe differences regarding the communication of different news topics as compared with ExxonMobil. For instance, filings related to *dividend payment* and *production outlook* feature a highly pessimistic tone and a risk level at the upper end of the scale. This is interesting, since Transocean Ltd. was implicated in the Deepwater Horizon oil spill resulting from the explosion of one of its oil rigs in the Gulf of Mexico. We also note that our findings match the actual valuation of Transocean Ltd. on the stock market. The stock price during our study period slipped from an all-time high of $160 in 2008 to an all-time low of $7.55 in 2017.

Figure 5: Risk-optimism matrix for news filings from Transocean Ltd.

5.4. Strategic analysis of the external environment

Our strategic framework enables managers to compare their firm with direct competitors. Figure 6 presents an instance in which we analyze all press releases from example energy firms with respect to *mergers and acquisitions* and *dividend payment*. Evidently, the strategic dimensions differ substantially across different companies. For instance, filings from ExxonMobil referring to *dividend payment* evince a highly optimistic tone with low risk, whereas those from Transocean Ltd. convey a more pessimistic tone with a risk level above average. As previously mentioned, this matches the actual development of these companies' stock market performance.

Figure 6: Risk-optimism matrix for filings related to *mergers and acquisitions* and *dividend payment*.

Figure 7 provides a strategic analysis for the sector as a whole. Here we

draw on the results from Table 1 (i.e., all filings in our dataset) and translate them into a highly interpretable visualization that resembles a SWOT analysis. For space reasons, we focus on the eight most frequent topics in our dataset, while we additionally include *legal issues* and *resource development* because of their particular relevance in the light of increasing policy changes and regulations. As previously mentioned, we see that the tone in the filings is highly dependent on the underlying topic. For example, filings related to *management change* typically feature a high level of optimism and convey high risk. At the other end of the scale, disclosures related to *legal issues* feature a low risk level together with a low optimism score. We also observe multiple topics in the center of the scales. For instance, filings related to *drilling contracts* are characterized by a medium optimism score in combination with a risk score above average.

Figure 7: Risk-optimism matrix for news topics from the energy sector.

As a major benefit, our approach also enables one to analyze the changes in tone of individual news topics over time. This is of particular interest in light of intensified climate policy in the energy sector. Figure 8 provides a strategic analysis for two different periods. The white points refer to all filings within the period from 2004 to 2012, whereas the gray points indicate filings published between 2013 and 2017. Evidently, the tone in the filings has trended toward lower optimism and higher risk scores. For example, the tone of *dividend payment* has shifted to a higher risk score, with a slightly lower optimism score. Similarly, filings referring to *production outlook* feature higher risk scores and lower optimism scores in the later period. The

corresponding difference between the optimism and risk scores is statistically significant at the 1 % significance level when a two-sided Welch t test is performed. Overall, we see that the interpretation of topics greatly depends on the time frame, and hence their connotation cannot be assumed to be fixed. Our approach provides a promising opportunity to overcome the drawbacks of current approaches by including a time dependency to study the dynamic reception of news topics.

Figure 8: Risk-optimism matrix for two different periods. The white points refer to all filings within the period from 2004 to 2012, whereas the gray points indicate filings published between 2013 and 2017.

5.5. Market response to risk and optimism

We now investigate the stock market reaction to the risk and optimism scores. For this purpose, we use an event study design to analyze the information value of a filing [57]. This allows us to estimate the effect of a filing on the stock market without confounding influences from the market itself. The cornerstone of this method is the assumption that stock prices in efficient markets reflect all information available to the market participants [58]. Granting this, we estimate a normal return in the absence of a filing and compare it with the observed return. The difference yields the abnormal return, which can be attributed to the novel information from the filing entering the market. We estimate the normal return according to the market model [57], which assumes a stable linear relationship between the market return and the normal return. The market return is modeled with use of the NYSE Composite Index, along with an event window of 10 trading days before the event.

We begin our stock market analysis by performing an analysis of variance test to examine whether there are any significant differences regarding the stock market effect across the identified topics. This is done by our partitioning the total variance of the abnormal returns into a component that is the result of true random error and the components that are attributable to the difference between the means of the topics. We observe an F statistic of 6.696, which is statistically significant at the 1 % significance level. Thus, we reject the null hypothesis of there being no difference between means, and conclude that the topics differ significantly in terms of the resulting stock market effect.

Next, we aim at analyzing the stock market reaction to risk and optimism for those topics that are particularly relevant to real-world business applications. To systematize the selection of these topics, we again performed an experiment with five students with a specialization in finance. We presented them with 10 random filings from each of the 20 topics in our dataset and asked which topics they perceive as particularly helpful. The majority of the students agreed on a list of 10 topics with an interrater reliability in terms of Fleiss's kappa of 0.589. Figure 9 visualizes these topics and analyzes their relationship to the average stock market reaction. To expose the role of the strategic dimensions, the diagram groups the filings from each topic into the following two categories: (a) filings with high optimism and low risk scores, which are shown in white; (b) filings with low optimism and high risk scores, which are shown in black. High risk (optimism) scores refer to filings with a risk (optimism) score above average within a given topic, whereas low risk (optimism) scores refer to filings with a risk (optimism) score below average.

Figure 9: Stock market reaction for different topics, but each separated for (a) filings with high optimism and low risk scores (white) and (b) filings with low optimism and high risk scores (black). High risk (optimism) scores refer to filings with a risk (optimism) score above average within a given topic, whereas low risk (optimism) scores refer to filings with a risk (optimism) score below average.

According to Figure 9, the reaction to risk and optimism cannot be assumed to be uniform; instead, it differs considerably across topics. For instance, the topic *earnings results* yields positive stock market reactions on average, while the difference between filings with low and high risk is relatively low. In contrast, we see that filings related to *production outlook* and *resource development* give a more differentiated picture. Here disclosures with high optimism and low risk result in positive abnormal returns, whereas filings with low optimism and high risk point in the opposite direction. We observe a different pattern for *management change*, whereby filings with high optimism and low risk are perceived more negatively than filings with low optimism and high risk. A possible explanation for this finding is that filings from this topic typically additionally refer to the current state of the company. As a result, investors might valuate a management change more positively if the company is in a critical state, and vice versa.

5.6. Comparison with the literature

We now detail four illustrative observations and link them to related research.

First, we find a noticeable coverage of issues related to mergers and acquisitions. This is not surprising given that the energy sector has been forced to adjust to the massive transformations brought about by technology, globalization, and policy changes. This also matches the observations of Kumar

et al. [59], who note that the sector is witnessing an increasing amount of forward integration and horizontal mergers. According to Dealogic statistics, the year 2011 alone saw mergers in the energy sector with an approximate value of $322 billion. Furthermore, in the light of a declining trend of oil prices, the need for efficiency of capital will be even more important and "another big wave of M&A activity in the oil and gas industry could soon break" [60].

Second, related research [61] also identifies a wealth of topics related to financial data, and yet measuring the linguistic style was not part of the work. Here our tool suggests high optimism and low risk components are associated with earnings. For instance, the second most common topic, *earnings results*, entails an average risk score of -0.80 and an optimism level of 0.60. This is in line with the recent influx of new investments in the sector, especially concerning renewable energy. In 2015, global investments in green energy reached $286 billion, which represented a sixfold increase in comparison with 2004; in contrast, half as much was invested in new gas and coal generation.[7] Most investment was allocated to asset financing (i.e., internally financing a company, debt, or equity), which explains the relatively low risk involved. However, there is definitely a certain amount of risk of running an unprofitable business, since the energy sector is known to have long amortizations. In this context, the sector is experiencing a declining trend in the self-financeable growth rate, which serves as an indicator for

[7]Global Trends in Renewable Energy Investment 2016: `http://fs-unep-centre.org/publications/global-trends-renewable-energy-investment-2016`, accessed October 10, 2017.

business viability and growth prospects [62].

Third, the topic *management change* entails a high level of optimism (i.e., 0.59). Related research across industries has already established the importance of announcements dealing with management changes [63]. The findings of previous studies yielded mixed outcomes regarding the direction in which stock prices move based purely on the presence of such a filing, while others explicitly identified a neutral abnormal return on the day of disclosure [37]. However, previous work neglected the actual content of the filing and, in contrast to our research, cannot sense the positive outlook.

Fourth, we note a considerable risk regarding *production outlook*, amounting to a risk score of 0.73. Sadorsky [14] found that energy-related stocks are twice as risky as the market benchmark. In this study, a 1 % movement in the underlying market index was associated with a 2 % movement in the stock price both upward and downward. A potential reason discussed in the literature is the considerable impact of policy changes in this sector. For instance, policy instruments such as feed-in tariffs for renewable energy can quickly change market dynamics and rule out forms of energy generation that were previously profitable. Accordingly, expectations of (sudden) policy changes affect the perceived risk of whether operations will remain sustainable. In the future, this might eventually result in a higher cost of capital. At the same time, energy policy is also regarded as a market barrier.

6. Discussion

6.1. Contribution

Research in the field of strategic management has a long tradition of devising management tools and concepts for measuring the performance of firms. Accordingly, our computational procedure contributes in the following aspects.

Our method builds on advanced analytics and thus benefits from being automated. Once implemented, it executes all computations in a fully computerized fashion, and accordingly, managers can update their performance assessments with arbitrary frequency. This is in contrast to common management tools such as industry reports from specialized agencies, firm-specific SWOT analyses, or growth-strength matrices. All of these require extensive manual labor, and are thus published only monthly or even less often, thereby running the risk of overlooking short-term trends that require immediate action.

Management frameworks predominantly address overall strengths and weaknesses on a highly abstract level. For instance, a growth-share matrix ranks different market segments or, less often, even products, while a SWOT analysis identifies common competences and issues prone to affect the enterprise as a whole or in large part. Unlike previous frameworks, ours actively engages in granular recommendations at the level of individual business units, activities, and processes.

Our text-based framework affords the opportunity to conduct holistic studies in the sense that the business units, activities, and processes subject to risk need not be known ex ante. This differs from common management

frameworks whereby the management has to define which items should be ranked a priori, thus entailing the definite possibility of failing to include relevant items because of various biases [64, 65]. As a remedy, our approach essentially draws on the *complete* knowledge encoded in narrative contents. It specifically performs an agnostic analysis in which the underlying themes are not predetermined but are instead extracted from the language.[8]

6.2. Managerial implications

This work specifically demonstrates how advanced analytics can provide business value. Thereby, it aids firms in strengthening their position in today's fierce market by competing with the help of analytics. This is in line with earlier claims by Gupta and George [66]. Some even argue in favor of a "fusion between IT strategy and business strategy," with big data being one element [67].

Our decision to computerize the SWOT analysis yields immediate business value to practitioners and businesses, which stems from the prevalence of SWOT analysis [5] as a vehicle in management consulting and strategic planning. According to a quote from Michael Watkins in the *Harvard Business Review*, "more than three-fourths of the participants in the executive

[8]This capability stems from the fact that our approach builds on topic modeling, whereby we can vary the number of different topic clusters. Depending on the managerial needs, practitioners can increase the number of topics so as to evaluate individual processes or activities. Conversely, one could reduce the number of topics to a handful, grouping documents by business unit instead of individual process. For instance, we performed our empirical demonstration with 20 topics as determined by an experiment. When applied to five topics, our method identifies more general themes that relate to different industry units. We thus used the experiments to come up with a labeling for this model initialization, which resulted in the following subjects: *shareholder rights*, *accounting*, *operations*, *legal*, and *investment*. Hence, the number of topics determines the depth of the strategic analysis in our method.

programs" use this type of analysis as a management tool. Beyond that, it is routinely adopted by organizations, as well as governments, and its prominence is undisputed by academic research according to a recent review on strategy tools in use [4]. This matches common practice in top-tier management consulting firms.

Moreover, our work presents a compelling case for the use of advanced analytics in another aspect: information systems research predominantly develops methods and tools for creating business value by mining *user data* (e.g., [65, 68, 69]), while we shift the focus toward leveraging *company data*. The use of firm-related data appears underrepresented in current research on big data analytics as outlined in recent review articles [8, 70]. Hence, as a managerial implication, we note that practitioners should carefully ponder further cases in which firm data can create value.

6.3. Limitations

Our approach is not free of limitations, even though it overcomes many of the shortcomings inherent in manual strategic frameworks. Linguistic content frequently entails noise because of its imprecision [13]. Language noise is especially prevalent in financial markets, where managers face an incentive to frame their disclosures in a certain way [47]. For instance, they often replace negative expressions with positive statements incorporating an additional negation term. In addition, behavioral research has shown that text mining can only approximate the subjective opinion of authors [40, 41], while a perfect translation of lingustic materials into numerical ratings seems out of reach because of behavioral constraints. Nevertheless, this presents an intriguing field of future research. Furthermore, our approach can only

sense the firm's performance as encoded in the narrative materials. Hence, special care is needed to circumvent any limitations that arise from the chosen source of news. A possible alternative is to combine different news sources: firm disclosures, such as press releases or regulatory filings, convey internal information, while newspaper articles, user reviews, etc., provide an external view.

7. Conclusion

Management tools have found widespread application in strategic planning. Among them is SWOT analysis, which, despite its age, is a vital tool, and any effort to automate its analysis is a source of direct value for firms and organizations. A potential computerization is reflected in our approach, which leverages recent innovations in advanced analytics and especially text mining. More specifically, it infers issues in entrepreneurial undertakings from narrative materials and assigns different risk-strength scores to these materials on the basis of the linguistic style. On the one hand, this allows one to track internal performance in core areas and it also functions as an early warning mechanism for critical developments. On the other hand, it elicits external comparisons with both competitors and the market environment in general. The inherent benefits are manifold, including automation, reproducible computation schemes, and differing levels of granularity, ranging from industry units to individual processes and activities. Our performance assessments can then help managers, who can then—in a subsequent step—devise and adapt their strategies accordingly. As such, our framework presents a highly flexible approach that effortlessly generalizes to arbitrary

firms, industries, and subjects.

Appendix A. Dictionary with risk and optimism/pessimism expressions

Our approach counts the frequencies of predefined words as follows.

Optimism/pessimism. We rely on the Loughran-McDonald dictionary to determine optimistic and pessimistic terms [47]. This word list was specifically designed with the characteristics of financial language in mind, and has found a wide range of applications in research [13]. We also experimented with alternative dictionaries, such as the Harvard IV psychological dictionary. Consistent with the previous literature, we find that these dictionaries yield similar results.

Risk. We are not aware of a dictionary specific to financial risk, and thus we constructed a dictionary that is tailored to such phrases. For this purpose, we asked five students with a specialization in finance to provide lists of words they regard as signaling risk. The process resulted in a list of 91 risk words that were labeled coherently by a majority of at least three students. We extended this list with risk-related expressions from the uncertainty list in the Loughran-McDonald dictionary.

Altogether, this resulted in 145 optimistic and 878 pessimistic terms, as well as a new dictionary of risk expression with 175 entries. The following list details all words labeled as expressions conveying risk. The list shows stems rather than complete words, because of stemming being part of the preprocessing.

```
abandon, abey, adjourn, adver, aggress, almost, alter,
ambigu, anomal, anticip, antitrust, appear, arbitrari,
assum, backlog, believ, borrow, break, breakag, catastroph,
caution, certain, chanceri, chase, claim, clarif, collap,
conceiv, condit, confus, conting, could, crude, damag,
deadlin, death, debt, declin, deep, defect, defer, deficit,
degr, depend, depreci, destabil, deterior, deviat, die,
differ, disabl, disagr, disclaim, discontinu, dismiss,
disrupt, divestitur, divid, doubt, downtim, downturn,
dramat, drop, exclud, exit, expenditur, exploit, exposur,
extrem, fluctuat, forfeit, heavi, hidden, hypothet,
imprecis, improb, incomplet, incorrect, indefinit,
indetermin, inexact, inflat, instabl, intang, loss, low,
may, might, minor, nonassess, occasion, ordinarili, pend,
perhap, possibl, precaut, preliminari, prepay, pressur,
presum, probabl, problem, ramp, random, reassess, recalcul,
recess, reconsid, reexamin, reinterpret, resign, revis,
risk, rough, rumor, seem, seldom, shut, sometim, somewhat,
somewher, specul, spent, sporad, sudden, suscept, tend,
tentat, terror, turbul, uncertain, unclear, unconfirm,
unconv, undecid, undefin, undesign, undetect, undetermin,
undocu, unexpect, unfamiliar, unfavor, unforecast,
unforeseen, unguarante, unhedg, unidentifi, uninsur,
unknown, unobserv, unplan, unpredict, unprov, unquantifi,
unreal, unreconcil, unregist, unschedul, unseason, unsettl,
unspecif, unsuccess, untest, untru, unwritten, urg, vagari,
vari, varianc, variant, violat, volatil, war, weaker
```

Appendix B. List of extracted topics

The following list shows the 20 most frequent word stems for each of the 20 extracted topics. In addition, we assign a short name to each topic, defining the overall subject. For example, *mergers and acquisitions* is the corresponding subject of topic 17.

Topic 1: Loan arrangement. *lender, borrow, loan, agent, administr, credit, bank, shall, commit, letter, revolv, rate, amount, document, agreement, assign, parti, default, day, hereund*

Topic 2: Trust indenture. *note, indentur, truste, holder, guarantor, redempt, restrict,*

interest, shall, payment, global, issuer, indebted, guarante, amount, supplement, default, transfer, person, offer

Topic 3: Legal issues. *prospectus, underwrit, offer, indemnifi, agreement, packag, therein, counsel, preliminari, free, respect, untru, purchas, write, sell, supplement, parti, deliveri, sale, opinion*

Topic 4: Earnings results. *quarter, incom, million, oper, net, earn, revenu, segment, compar, per, dilut, cash, loss, expens, total, adjust, month, increas, measur, tax*

Topic 5: Income statements. *gas, oil, per, natur, product, net, quarter, averag, price, million, oper, total, boe, deriv, volum, cash, incom, hedg, expens, realiz*

Topic 6: Security agreement. *subsidiari, lien, shall, interest, payment, properti, respect, debtor, collater, restrict, claim, bankruptci, transact, case, reason, amount, permit, document, agent, thereof*

Topic 7: Employment agreement. *agreement, parti, shall, servic, group, termin, agre, arbitr, term, reason, confidenti, entiti, law, notic, right, claim, breach, employe, indemnif, lesse*

Topic 8: Purchase agreement. *seller, buyer, close, purchas, parti, agreement, defect, shall, asset, properti, warranti, schedul, respect, represent, indemnifi, transact, tax, claim, contempl, right*

Topic 9: Tax report. *decemb, solid, net, asset, cash, incom, cost, total, consolid, forma, fair, tax, loss, oper, pro, million, deriv, condens, adjust, expens*

Topic 10: Stock option award. *plan, award, employe, particip, committe, grant, vest, termin, incent, shall, compens, restrict, bonus, salari, payment, disabl, determin, period, agreement, death*

Topic 11: Resource development. *well, drill, reserv, develop, amp, product, gas, shale, acr, oil, basin, field, play, prove, acreag, explor, resourc, reservoir, horizont, eagl*

Topic 12: Management change. *director, meet, board, stockhold, corpor, vote, proxi, elect, shall, bylaw, nomin, chairman, committe, secretari, person, sharehold, propos, notic, nomine, appoint*

Topic 13: Amendment of shareholder rights. *seri, holder, right, shall, prefer, class, distribut, member, certif, convers, transfer, person, respect, determin, alloc, notic,*

vote, upon, entitl, adjust

Topic 14: Oil price development. *fuel, refin, refineri, crude, project, product, market, growth, barrel, retail, margin, gasolin, billion, sourc, chemic, million, oper, export, demand, improv*

Topic 15: Infrastructure and logistics. *pipelin, logist, storag, distribut, crude, partner, refin, transport, throughput, termin, system, oper, facil, refineri, volum, asset, mainten, plain, acquisit, tank*

Topic 16: Partnership arrangement. *partnership, partner, limit, interest, entiti, agreement, common, plain, contribut, membership, hold, distribut, alloc, member, amend, oper, subordin, approv, incent, midstream*

Topic 17: Mergers and acquisitions. *amend, parent, agreement, merger, effect, restat, herebi, page, parti, second, credit, inc, transact, first, document, none, sec, energi, represent, delet*

Topic 18: Public relations. *coal, releas, trust, energi, press, royalti, mine, ton, inc, messag, news, distribut, april, august, octob, januari, juli, march, novemb, februari*

Topic 19: Dividend payment. *share, stock, common, dividend, option, convert, exercis, sharehold, prefer, price, corpor, convers, par, stockhold, fundament, transact, offer, split, repurchas, close*

Topic 20: Drilling contracts. *rig, contract, mid, drill, risk, offshor, fleet, late, earli, low, mexico, status, mar, gulf, dec, plus, day, sea, hurrican, upgrad*

References

[1] G. G. Dess, A. B. Eisner, G. T. Lumpkin, Strategic management: Text and cases, McGraw-Hill/Irwin, New York, NY, 4 edition, 2008.

[2] B. de Wit, R. Meyer, Strategy: Process, content, context, Thomson, London, UK, 3 edition, 2009.

[3] C. Cheng, M. I. Havenvid, Investigating strategy tools from an interactive perspective, IMP Journal 11 (2017) 127–149.

[4] P. Jarzabkowski, S. Kaplan, Strategy tools-in-use: A framework for understanding "technologies of rationality" in practice, Strategic Management Journal 36 (2015) 537–558.

[5] S. Ghazinoory, M. Abdi, M. Azadegan-Mehr, SWOT methodology: A state-of-the-art review for the past, a framework for the future, Journal of Business Economics and Management 12 (2011) 24–48.

[6] S. A. Zahra, S. S. Chaples, Blind spots in competitive analysis, Academy of Management Executive 7 (1993) 7–28.

[7] J. W. Alba, J. W. Hutchinson, Dimensions of consumer expertise, Journal of Consumer Research 13 (1987) 411–454.

[8] P. Mikalef, I. O. Pappas, J. Krogstie, M. Giannakos, Big data analytics capabilities: A systematic literature review and research agenda, Information Systems and e-Business Management forthcoming (2017).

[9] S. Elliot, Transdisciplinary perspectives on environmental sustainability: A resource base and framework for IT-enabled business transformation, MIS Quarterly 35 (2011) 197–236.

[10] J. vom Brocke, R. T. Watson, C. Dwyer, S. Elliot, N. Melville, Green information systems: Directives for the IS discipline, Communications of the Association for Information Systems 33 (2013) 509–520.

[11] R. J. Baumgartner, R. Rauter, Strategic perspectives of corporate sustainability management to develop a sustainable organization, Journal of Cleaner Production 140 (2017) 81–92.

[12] R. G. Eccles, K. M. Perkins, G. Serafeim, How to become a sustainable company, MIT Sloan Management 53 (2012) 43–50.

[13] T. I. Loughran, B. McDonald, Textual analysis in accounting and finance: A survey, Journal of Accounting Research 54 (2016) 1187–1230.

[14] P. Sadorsky, Modeling renewable energy company risk, Energy Policy 40 (2012) 39–48.

[15] J. Li, X. Li, B. Zhu, User opinion classification in social media: A global consistency maximization approach, Information & Management 53 (2016) 987–996.

[16] K. Coussement, D. van den Poel, Integrating the voice of customers through call center emails into a decision support system for churn prediction, Information & Management 45 (2008) 164–174.

[17] M. Eickhoff, J. Muntermann, Stock analysts vs. the crowd: Mutual prediction and the drivers of crowd wisdom, Information & Management (2016).

[18] S. Feuerriegel, H. Prendinger, News-based trading strategies, Decision Support Systems 90 (2016) 65–74.

[19] M. Kraus, S. Feuerriegel, Decision support from financial disclosures with deep neural networks and transfer learning, Decision Support Systems forthcoming (2017).

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

[21] N. Goby, T. Brandt, S. Feuerriegel, D. Neumann, Business intelligence for business processes: The case of IT incident management, in: European Conference on Information Systems (ECIS).

[22] N. Pröllochs, S. Feuerriegel, D. Neumann, Negation scope detection in sentiment analysis: Decision support for news-driven trading, Decision Support Systems 88 (2016) 67–75.

[23] Y. Rao, H. Xie, J. Li, F. Jin, F. L. Wang, Q. Li, Social emotion classification of short text via topic-level maximum entropy model, Information & Management 53 (2016) 978–986.

[24] Z. Shi, G. Lee, A. B. Whinston, Toward a better measure of business proximity: Topic modeling for industry intelligence, MIS Quarterly 40 (2016) 1035–1056.

[25] T. Brandt, J. Bendler, D. Neumann, Social media analytics and value creation in urban smart tourism ecosystems, Information & Management 54 (2017) 703–713.

[26] P. Bromiley, D. Rau, Towards a practice-based view of strategy, Strategic Management Journal 35 (2014) 1249–1256.

[27] N. Capon, J. U. Farley, J. M. Hulbert, Strategic planning and financial performance: More evidence, Journal of Management Studies 31 (1994) 105–110.

[28] E. Vaara, J.-A. Lamberg, Taking historical embeddedness seriously: Three historical approaches to advance strategy process and practice research, Academy of Management Review 41 (2016) 633–657.

[29] T. M. Amabile, A model of creativity and innovation in organizations, Research in Organizational Behavior 10 (1988) 123–167.

[30] S. G. Bharadwaj, P. R. Varadarajan, J. Fahy, Sustainable competitive advantage in service industries: A conceptual model and research propositions, Journal of Marketing 57 (1993) 83–99.

[31] B. Wooldridge, S. W. Floyd, Research notes and communications strategic process effects on consensus, Strategic Management Journal 10 (1989) 295–302.

[32] C. A. O'Reilly, D. F. Caldwell, J. A. Chatman, M. Lapiz, W. Self, How leadership matters: The effects of leaders' alignment on strategy implementation, The Leadership Quarterly 21 (2010) 104–113.

[33] T. C. Powell, D. Lovallo, C. R. Fox, Behavioral strategy, Strategic Management Journal 32 (2011) 1369–1386.

[34] C. M. Bishop, Pattern recognition and machine learning, Springer, New York, NY, 11 edition, 2013.

[35] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, The Journal of Machine Learning Research 3 (2003) 993–1022.

[36] D. M. Blei, Probabilistic topic models, Communications of the ACM 55 (2012) 77–84.

[37] S. Feuerriegel, A. Ratku, D. Neumann, Analysis of how underlying topics in financial news affect stock prices using latent Dirichlet allocation, in: 49th Hawaii International Conference on System Sciences (HICSS), IEEE Computer Society, 2016, pp. 1072–1081.

[38] P. C. Tetlock, M. Saar-Tsechansky, S. Macskassy, More than words: Quantifying language to measure firms' fundamentals, Journal of Finance 63 (2008) 1437–1467.

[39] K. W. Hanley, G. Hoberg, Litigation risk, strategic disclosure and the underpricing of initial public offerings, Journal of Financial Economics 103 (2012) 235–254.

[40] R. Feldman, Techniques and applications for sentiment analysis, Communications of the ACM 56 (2013) 82–89.

[41] B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval 2 (2008) 1–135.

[42] K. Ravi, V. Ravi, A survey on opinion mining and sentiment analysis: Tasks, approaches and applications, Knowledge-Based Systems 89 (2015) 14–46.

[43] C. Kearney, S. Liu, Textual sentiment in finance: A survey of methods and models, International Review of Financial Analysis 33 (2014) 171–185.

[44] F. Li, Textual analysis of corporate disclosures: A survey of the literature, Journal of Accounting Literature 29 (2010) 143–165.

[45] C. D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA, 1999.

[46] A. Yadollahi, A. G. Shahraki, O. R. Zaiane, Current state of text sentiment analysis from opinion to emotion mining, ACM Computing Surveys 50 (2017) 1–33.

[47] T. Loughran, B. McDonald, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, Journal of Finance 66 (2011) 35–65.

[48] M. Steyvers, T. Griffiths, Probabilistic topic models, in: T. K. Landauer (Ed.), Handbook of Latent Semantic Analysis, Psychology Press, Mahwah, NJ, 2013, pp. 424–440.

[49] J. Chang, J. Boyd-Graber, C. Wang, S. Gerrish, D. M. Blei, Reading tea leaves: How humans interpret topic models, in: Advances in Neural Information Processing Systems (NIPS), pp. 288–296.

[50] C. Sievert, K. E. Shirley, Ldavis: a method for visualizing and interpreting topics, in: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pp. 63–70.

[51] N. Pröllochs, S. Feuerriegel, D. Neumann, Generating domain-specific dictionaries using bayesian learning, in: 23rd European Conference on Information Systems (ECIS).

[52] S. Feuerriegel, R. Fehrer, Improving decision analytics with deep learning: The case of financial disclosures, in: 24th European Conference on Information Systems (ECIS).

[53] M. J. Bürer, R. Wüstenhagen, Which renewable energy policy is a venture capitalist's best friend? Empirical evidence from a survey of international cleantech investors, Energy Policy 37 (2009) 4997–5006.

[54] D. Ramage, D. Hall, R. Nallapati, C. D. Manning, 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, EMNLP '09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 248–256.

[55] V. Niederhoffer, The analysis of world events and stock prices, Journal of Business 44 (1971) 193–219.

[56] M. E. Carter, B. S. Soo, The relevance of Form 8-K reports, Journal of Accounting Research 37 (1999) 119.

[57] A. C. MacKinlay, Event studies in economics and finance, Journal of Economic Literature 35 (1997) 13–39.

[58] Y. Konchitchki, D. E. O'Leary, Event study methodologies in information systems research, International Journal of Accounting Information Systems 12 (2011) 99–115.

[59] S. Kumar, S. Managi, A. Matsuda, Stock prices of clean energy firms, oil and carbon markets: A vector autoregressive analysis, Energy Economics 34 (2012) 215–226.

[60] B. Evans, S. Nyquist, K. Yanosek, et al., Mergers in a low oil-price environment: Proceed with caution, Journal of Petroleum Technology 68 (2016) 49–51.

[61] N. Székely, J. vom Brocke, What can we learn from corporate sustainability reporting? deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique, PLoS One 12 (2017) e0174807.

[62] P. D. Lund, How fast can businesses in the new energy sector grow? an analysis of critical factors, Renewable Energy 66 (2014) 33–40.

[63] A. Neuhierl, A. Scherbina, B. Schlusche, Market reaction to corporate press releases, Journal of Financial and Quantitative Analysis 48 (2013) 1207–1240.

[64] D. Galai, O. Sade, The "ostrich effect" and the relationship between the liquidity and the yields of financial assets, Journal of Business 79 (2006) 2741–2759.

[65] N. Hu, P. A. Pavlou, J. Zhang, On self-selection biases in online product reviews, MIS Quarterly 41 (2017) 449–471.

[66] M. Gupta, J. F. George, Toward the development of a big data analytics capability, Information & Management 53 (2016) 1049–1064.

[67] A. Bharadwaj, O. A. El Sawy, P. A. Pavlou, N. Venkatraman, Digital business strategy: Toward a next generation of insights, MIS Quarterly 37 (2013) 471–482.

[68] D. Martens, F. Provost, J. Clark, E. J. de Fortuny, Mining massive fine-grained behavior data to improve predictive analytics, MIS Quarterly 40 (2016) 869–888.

[69] J. Qi, Z. Zhang, S. Jeon, Y. Zhou, Mining customer requirements from online reviews: A product improvement perspective, Information & Management 53 (2016) 951–963.

[70] C. L. P. Chen, C.-Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences 275 (2014) 314–347.

Author biographies

Nicolas Pröllochs

Nicolas Pröllochs is a postdoctoral researcher at the Department of Engineering Science of the University of Oxford. Previously, he headed a research group in Social Computing at the University of Freiburg where he also obtained his Ph.D. His research focuses on computational techniques for understanding human decision-making in the digital age. Current research applies quantitative text analysis and natural language processing to a broad selection of topics, including financial markets, customer analysis, and recommendation systems. He has co-authored research publications at International Conference on Information Systems, Hawaii International Conference on System Sciences and Decision Support Systems.

Stefan Feuerriegel

Stefan Feuerriegel is an assistant professor for management information systems at ETH Zurich. His research focuses on cognitive information systems and business intelligence, including text mining and sentiment analysis of financial news. Previously, he obtained his Ph.D. from the University of Freiburg where also worked as a research group leader at the Chair for Information Systems Research. He has co-authored research publications in the European Journal of Operational Research, the European Journal of Information Systems, the Journal of Information Technology and Decision Support Systems.