Information Sciences Letters

Volume 12
Issue 5 *May 2023*

Article 18

2023

# A Business Analytics Approach to Strategic Management using Uncovering Corporate Challenges through Topic Modeling

A. Y. Nasereddin
*Faculty of Business Administration, Middle East University, Amman 11831, Jordan*,
nasereddin@meu.edu.jo

Information Sciences Letters
*An International Journal*

http://dx.doi.org/10.18576/isl/120518

# A Business Analytics Approach to Strategic Management using Uncovering Corporate Challenges through Topic Modeling

*A. Y. Nasereddin*

Faculty of Business Administration, Middle East University, Amman 11831, Jordan

Received: 21 Feb. 2022, Revised: 22 Mar. 2022, Accepted: 24 Mar. 2022.
Published online: 1 May 2023.

**Abstract:** Business analytics is a robust strategic management tool, and topic modeling is a technique that can be leveraged to derive insights from vast collections of unstructured data. Topic modeling is an automated method that identifies abstract concepts, or "topics," present in various data sources, such as customer feedback, social media posts, and news articles. Through topic modeling, organizations can gain a better understanding of their customers, competitors, and market trends, which can be used to make informed strategic decisions, such as identifying new business opportunities, enhancing product or service offerings, and recognizing potential risks. Moreover, by integrating topic modeling with other business analytics approaches, such as predictive modeling, organizations can gain a more comprehensive perspective of their performance and make data-driven decisions. In essence, topic modeling is a valuable tool for strategic management that provides organizations with the insights they need to stay ahead of the competition and make informed decisions. To make effective strategic decisions, it is crucial to comprehend an organization's internal and external environments fully. The proposed approach utilizes text-mining techniques to augment traditional management tools, such as SWOT analysis or growth-share matrix. By examining narrative materials, such as financial disclosures, we apply topic modeling to identify critical challenges faced by an organization. We then quantify the language used in these materials in terms of risk and optimism, which provides a detailed understanding of a company's strengths and weaknesses and helps identify business units, activities, and processes that may be at risk. Additionally, this approach can be used to compare a company with its competitors or the broader market.

**Keywords:** Business Approach; Text mining; Performance Analysis; Topic modeling; Strategic management.

## 1 Introduction

Business analytics is a set of techniques used to analyze data and make strategic decisions in an organization [1, 2]. It involves using various techniques such as statistical analysis, data mining, machine learning, and predictive modeling to understand and forecast business performance. Business analytics can be used for a wide range of strategic decisions, including forecasting sales, identifying new markets, optimizing operations, and managing risk [3, 4]. Additionally, it can be used for data-driven marketing, finance, and human resources decisions. Business analytics enables organizations to make better decisions, improve performance, and gain a competitive advantage by leveraging the vast amounts of data available to them [5, 6].

Business analytics for strategic management is a field that aims to help organizations make data-driven decisions by leveraging advanced analytics techniques [7]. One of the key challenges in strategic management is identifying and assessing corporate challenges. Traditional approaches such as SWOT analysis or growth-share matrix can be time-consuming and may not provide a comprehensive view of a company's internal and external environments [8]. Using a technique called topic modeling, it is possible to conclude huge amounts of unstructured data, including customer comments, posts on social media, and news items. Organizations may better understand their consumers, rivals, and market trends using topic modeling to various data sources. These can be used to make strategic choices like seeing new business possibilities, enhancing product or service offerings, or spotting possible hazards [9].

The method of making choices and taking actions that will influence and lead an organization's overall success is known as strategic management [10]. The ultimate goal of strategic management is for the business to establish and maintain a competitive edge. Understanding the organization's current condition is the first step in the ongoing process of strategic management. Next, looking forward to seeing potential dangers and opportunities that impact the organization's future is important. This process includes assessing the organization's assets and liabilities and external opportunities and risks. In this process, the organization's internal and external surroundings are analyzed, goals and

*Corresponding author e-mail: nasereddin@meu.edu.jo

objectives are determined, and strategies are developed to help them be met. Finally, the strategies are implemented, and the results are tracked. Based on this analysis, strategic managers develop long-term plans and strategies to achieve organizational goals [11, 12].

Effective strategic management requires collaboration and coordination among different levels of management and different functional areas within the organization. It also requires clear communication and alignment of goals and objectives throughout the organization. Strategic management is important for organizations of all sizes and types, as it helps them to stay competitive and adapt to changing conditions. It is also important for organizations to regularly engage in strategic management in order to stay ahead of the competition and achieve their goals in the long-term.

Strategic management via topic modeling is a process of using advanced analytics and natural language processing techniques to analyze narrative materials from an organization, such as financial disclosures, press releases, and regulatory filings. The objective is to identify the organization's main problems and difficulties and to measure the terminology used in those documents along two dimensions: risk and optimism. Topic modeling is a technique that uses machine learning algorithms to identify themes or topics in large collections of text data. It can identify latent themes that may not be immediately obvious and can be used to extract insights from unstructured data.

By applying topic modeling to narrative materials from an organization, strategic managers can gain a better understanding of the internal and external factors that may affect the organization. This can help them to identify strengths and weaknesses, and to develop strategies to address those challenges. Topic modeling can also be used to compare an organization's performance with that of competitors or the market in general. By analyzing the language utilization in relation to risk and optimism, strategic managers can identify patterns and trends that may not be immediately obvious. Overall, strategic management via topic modeling is a powerful tool for organizations to gain insights into their internal and external environments, identify and assess corporate challenges, and develop effective strategies to address them.

Optimism and risk are the two variables along which we evaluate language usage. In this study, we offer a methodology for text analysis mining that uses topic modeling to extract the main problems an organization faces from narrative resources like financial statements. This enables us to compare a firm to its rivals or the market and disclose its strengths and weaknesses by highlighting business units, activities, and procedures that are at risk. Our approach can help organizations to make data-driven decisions and stay ahead of the competition by identifying key issues in a timely and efficient manner.

The development of management tools and theories for gauging business success has long been a goal of strategy research. Below are some ways that our computational method advances this field:

Because our system is computerized and simple, managers often update their performance evaluations. This contrasts with conventional management tools like industry analysts, SWOT assessments, and growth-strength models, which need manual work and are often released only monthly or less frequently, increasing the danger of missing short-term patterns that call for quick action.

Conventional management frameworks frequently concentrate on general functioning effectively and shortcomings. Unlike other approaches, ours actively participates in detailed suggestions at the level of specific business units, operations, and processes.

Our text-based approach enables holistic analyses in that it is not necessary to know which business units, activities, and processes are at risk ahead. This contrasts with conventional management structures, which force managers to determine the items' rankings beforehand and incur the danger of excluding pertinent things due to various biases. Our method provides an agnostic analysis where the underlying concepts are not preconceived but are instead derived from the language since it is based on the complete information embedded in narrative materials.

The remainder of this article is provided as follows. The most comparable works are displayed in Section 2. The suggested framework is displayed in Section 3. The experiments and discussion are shown in Section 4. The conclusion and the next work directions are provided in Section 5.

## 2 Related works

Topic modeling is a method used in computer science that is gaining popularity in management research. It allows researchers to uncover the underlying themes and concepts in large amounts of textual data, such as documents or articles. By breaking down text into distinct topics, it helps researchers to understand the relationships between different ideas and identify patterns in the data. This study aims to demonstrate how topic modeling can be used in management research to advance our understanding of various phenomenon, by walking through the steps of the process and analyzing its application in management articles [13]. It demonstrates how topic modeling may be used to identify new

and developing concepts, create inductive categorization schemes, comprehend online audiences and goods, examine activist groups and frameworks, and comprehend cultural dynamics. The examination of topic modeling's most recent developments and the function of researcher interpretation in computer-driven research comes to a close.

Strategic technology planning is crucial for businesses to stay competitive in today's ever-changing market. One of the key tools used in this process is technology roadmaps, which link technologies to their respective markets. With the advent of big data analytics, there has been a growing interest in developing data-driven technology roadmaps. However, there is a lack of systematic methods for creating these roadmaps. To address this gap, this study proposes a framework for developing data-driven technology roadmaps [14]. The framework is comprised of three phases: layer mapping, content mapping, and opportunity finding. The first phase uses topic modeling to identify sub-layers for the technology roadmap. The second phase uses keyword network analysis for content mapping. Finally, the third phase uses link prediction to identify potential future innovations. This study provides a systematic method for creating data-driven technology roadmaps and offers data-driven evidence to aid in decision-making.

The field of IS research has undergone significant changes over time [15]. Using topic modeling with latent semantic indexing, author-supplied keywords were analyzed through the evaluation of 2962 papers published in prestigious IS journals between 2003-2017. These results show that despite keeping its basic purpose and character, IS research has expanded to include new fields. The study revealed that while the popularity of some issues, like e-commerce and IT outsourcing, has fluctuated over time, others, including IS innovation, IT deployment, and IS consumption, have stayed relatively constant. Newer subjects have also acquired popularity recently, including online communities, design science, and social media.

In this study [16], The authors use natural language processing, deep learning, and regression methods to analyze the current literature to evaluate AI's possibilities in strategic marketing. To map the existing body of knowledge, the writers seek to find recurrent themes, variety, historical progression, and dynamic elements in literature. The findings highlight ten major study issues, including comprehending consumer attitudes, utilizing AI to enhance market performance, and assessing customer contentment. Additionally, they provide a detailed analysis of key concepts, co-occurring keywords, authorship networks, and landmark publications in the field. Based on our findings, they propose a research agenda for future studies in the intersection of AI and strategic marketing.

This study proposed an automated text mining framework for analyzing the electronic document stream on supplier management platforms in the automotive industry [17]. In order to identify the profit position for making purchases and make a significant contribution to supply chain cost management, the approach includes classification techniques and descriptive analysis to analyze both textual elements (such as requests for data and offers) and narrative content (such as economic and computation data). The research illustrates the approach of using service provider documentation in buying procedures.

In this paper [18], they examine how big data analytics may be used in a marketing mix framework and how it might offer insights for more informed marketing decisions. They examine current research difficulties and future goals in big data analytics and target marketing. They highlight important data sources, methodologies, and techniques connected to five core marketing viewpoints: people, product, place, pricing, and marketing.

In the highly competitive manufacturing industry, the advent of the Fourth Industrial Revolution has led to increased research into the implementation and success of smart factories [19]. These factories are seen as a solution to complex manufacturing challenges and a way to achieve sustainable growth. However, as smart factory research spans multiple disciplines, it is important to understand past and current research trends to ensure their successful implementation. This study used topic modeling and regression-based methods to analyze the research trends of smart factories in international and Korean studies. The results revealed clear trends and allowed for comparison between the research trends in Korea and internationally. The findings and suggestions presented in this study can be used to guide future strategies for the diffusion of smart factories.

Topic modeling is a widely used method for identifying latent patterns in large data sets and has gained popularity in the field of marketing in recent years [20]. Despite the growing interest in using topic models in marketing research, there is currently no comprehensive overview of the field. There has been significant progress in various marketing sub-areas, however, there is still room for future research, particularly in integrating multiple dynamic data sources and combining exploratory topic models with predictive marketing models.

## 3 The Proposed Research Framework

Our computational procedures are first explained with an explanation of the intuition behind them, then human language preparation and topic-dependent word processing are covered.

### 3.1 Theoretical Foundation

The detection and assessment of company difficulties through topic modeling provide the theoretical basis for business intelligence in strategic management. Topic modeling is a method of uncovering latent patterns and structure in text data by identifying common themes or topics in a corpus of text. By analyzing a company's internal and external communications, such as emails, reports, and social media posts, topic modeling can reveal insights into the organization's challenges and opportunities. These insights can inform strategic decision-making by providing a deeper understanding of the company's environment and the factors that may impact its success.

This article focuses on practical techniques for developing strategies from a practice-based perspective, as opposed to the resource-based view, which focuses on unique activities that other firms cannot replicate [21]. The practice-based approach is more relevant in our setting as strategies are viewed as imitable activities or practices that can be transferred across firms.

Much research has been done on the results of successful strategies, including identifying factors that facilitate or hinder strategy implementation [22]. A major meta-analysis found that companies that use a formal planning process tend to perform better. But this requires a good match between the strategy and the business environment. Recent studies also examine creating a strategy, considering how it changes over time [23].

Situational analysis gathers information about a company's internal resources and external opportunities and challenges [24]. Comparing this data is intended to find measures that may be taken to enhance the firm's current status. SWOT analysis uses this as a management technique to give useful insights into the market and prospective areas for performance improvement [25].

Research on organizational learning shows a connection between a company's internal capabilities and its market performance [26]. Organizational learning is particularly effective when strategy and implementation are well-aligned. Organizational structure, specifically leadership, plays a key role in this process. For example, leadership alignment affects the successful implementation of strategies. Vaara and Lamberg studied the organizational structure processes necessary for successful strategic changes [27]. Behavioral aspects such as heuristic rules in decision-making can also affect the successful implementation of strategies, as discussed in [28].

### 3.2 Framework Used

Previous strategic management research has resulted in various tools, as outlined in [25]. According to Alba and Hutchinson [24], analyzing risks and opportunities is crucial for strategic planning, which is why our work focuses on operationalizing these dimensions through SWOT analysis. This choice is supported by several practical reasons, as discussed below.

SWOT analysis is a crucial technique for consulting businesses and is frequently used in performance management and company strategy. Recent studies support the fact that it is widely employed in many organizations. Examples include public organizations like UNICEF and governmental entities like the Local authorities and the European Commission. This is congruent with the methods used by top-tier strategy consulting companies, who frequently base their work on SWOT analysis. Given these factors, we decided to convert SWOT analysis into automated processes since this is most likely to assist firms and have an immediate influence on management practice.

### 3.3 Method Outline

We are developing computational procedures to track a company's performance throughout various business divisions, processes, and activities. This method, which includes topic identification and language analysis particular to each topic, is shown in broad strokes in Figure 2. The idea of language analysis has previously been used in strategic planning, where it is necessary to manually search for terms that point to rivals' cost-conscious conduct [29].

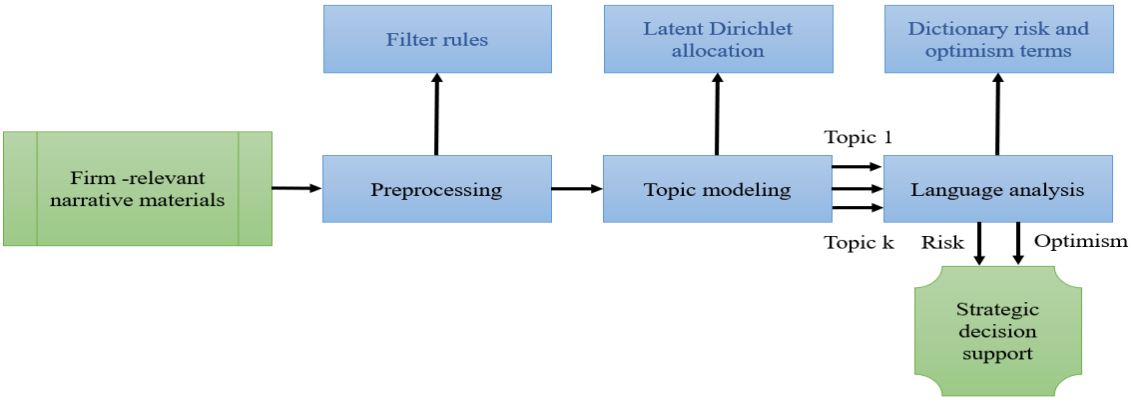


**Fig. 1.** The suggested strategic approach involves first identifying relevant topics, and then analyzing their language.

Our proposed strategic framework utilizes narrative materials, which sets it apart from traditional methods in strategic management such as surveys, expert opinions, and heuristic rules. Using natural language has several advantages: first, it allows for a comprehensive analysis of a company's strategic position and overcomes the potential challenges of managers strategically ignoring certain issues. Second, our methods offer a simple, quantifiable statistic to assess across enterprises. Third, our method can pick up on nuanced viewpoints like a company's individualized policy risk assessment.

The narrative materials can come from various sources, each with its own benefits. Using regulatory filings highlights a company's internal reporting and quantifies management's perspective on the company's performance. Media articles, on the other hand, focus on the perceived performance of companies and can detect issues that may not have been reported in regulatory filings.

Our system uses topic modeling to determine the various topics in the corpus [30]. We support topic modeling since it enables a very flexible level of analytical granularity and offers an automated mechanism for segmenting the corpus into topics. Unlike conventional approaches to strategic management, our method immediately infers company issues and dimensions from the facts rather than prescribing them. This enables us to learn about the numerous problems businesses are disclosing or being covered by the media.

After subject modeling, we assign ratings to each topic's files based on the risk and optimism aspects of a strategy following academic research and the intuition of the SWOT matrix. These factors, thought to influence financial market economic choices, are of special relevance to management since they direct how things are moved along these dimensions: While risk factors should be changed into non-risk factors, managers see a comparative benefit in high-performing goods.

The real risk and optimism ratings are carried out by scrutinizing the language nuances present in the papers. Finance has prioritized extracting positivity and gloom from exposures or news, and scholars have created special collections of language terms that suggest an optimistic or gloomy attitude [31, 32]. Similarly, businesses are known to include risk-related data in narrative documents to reduce the potential of lawsuits, particularly during the initial public offerings [33]. The widely held belief is that these methods can accurately and reliably deduce the (subjective) data and convert it into scales for the intended use. This presumption has constantly been proven correct, especially in contexts related to accounting. Common rule-based methods for measuring the frequencies of word labels depending on the prior characteristics [34].

### 3.4 Preprocessing

Text mining commonly involves several operations such as text cleaning, tokenization, stemming/lemmatization, stop word removal, and feature extraction/selection [35, 36]. After these operations are completed, the resulting narrative materials can be used for further analysis such as sentiment analysis, topic modeling, and text classification [37, 38]. These materials can also be used to train machine learning models for natural language processing tasks.

In the first step of our process, we preprocess narrative materials using standard text mining techniques, such as text cleaning, tokenization, stemming/lemmatization, stop word removal, and feature extraction/selection. This preparation is done to structure the text in a way that allows for further analysis. We remove irrelevant information such as contact addresses and formatting, keep only meaningful words and truncate inflected words to their stem, and remove words that appear in less than 1% of the documents to reduce skewness in the data [39, 40]. After the initial data preparation, t refers to any individual word in the corpus and tft refers to the number of times that word appears in the corpus. The

resulting word frequencies are used for further analysis.

### 3.5 Topic Modeling

Topic modeling is a technique used in business analytics for strategic management to identify the main topics or themes in a collection of documents [7, 16]. This technique is based on natural language processing and machine learning algorithms, which analyze the text and identify patterns in the data. The goal of topic modeling is to find latent structures in the data that reveal the underlying topics or themes that are discussed in the documents. These topics can then be used to gain insights and make strategic decisions in various business areas such as market research, customer segmentation, and product development.

Topic modeling is a statistical method used to identify themes or topics within a collection of documents [30, 41]. This technique uses word frequencies to group documents into clusters of similar content. LDA is a popular probabilistic model used in topic modeling. This method's ability to make use of effective probability inference algorithms and provide understandable subjects unsupervised is one of its primary features. This procedure relies on the belief that a certain collection of subjects produces each document. Each subject is treated mathematically as a distribution of vocabulary words, and each document is represented as a dispersion of subjects.

- **Document-topic Relationship**

The document-topic relationship in topic modeling refers to the way in which each document is associated with one or more topics. This relationship is typically represented mathematically by a probability distribution, where the probability of a document being associated with a particular topic is proportional to the relevance of that topic to the document. This relationship is established using algorithms such as Latent Dirichlet Allocation (LDA) which analyze the text in the document and assign probabilities to the different topics based on the words in the document. These probabilities can then be used to identify the main topics discussed in the document and how they are related to other documents in the corpus.

The mathematical notion of the document-topic relationship in topic modeling is typically represented by a matrix known as the "document-topic matrix". This matrix is created by applying a topic modeling algorithm such as Latent Dirichlet Allocation (LDA) to a corpus of documents. Each row of the matrix represents a document in the corpus, and each column represents a topic. The entries of the matrix are the probabilities of a document belonging to a particular topic, which are calculated by analyzing the text in the document and determining the relevance of each topic to the document based on the words it contains. In more detail, the mathematical representation of document-topic relationship is defined as follows:

Given a corpus of D documents and K topics, the document-topic relationship is represented by a D x K matrix, denoted by $\theta$, where $\theta_{d,k}$ is the probability of topic k in document d. In other words, $\theta_{d,k}$ represents the probability of topic k generating the words in document d. The entries in this matrix are non-negative and sum to one.

This matrix can be used to identify the main topics discussed in each document and how they are related to other documents in the corpus. It can also be used to identify the specific topics that are most strongly associated with a document, which can be useful for text summarization, keyword extraction, and text classification.

- **Word Frequency**

Word frequency refers to the number of times a specific word appears in a text or a collection of texts (corpus). It is a measure of how often a word is used in a text. This information can be used in various NLP tasks such as text mining, natural language understanding, and text generation. Word frequency can be used to identify the most important or relevant words in a text, which can be useful for text summarization, keyword extraction, and text classification. Additionally, it can also help to understand the general theme or topic of a text. High frequency words that occur in a specific context can be used to identify the topic of a text. Word frequency is often used in combination with other features such as word co-occurrence, part-of-speech tagging, and syntactic parsing to extract more meaningful information from a text.

The mathematical notion of word frequency in a text or corpus of texts is typically represented by a vector known as the "term frequency vector" or "word frequency vector". This vector is created by analyzing the text and counting the number of occurrences of each word or term in the text. Each entry in the vector represents the frequency of a specific word or term in the text.

In more detail, the mathematical representation of word frequency is defined as follows:

Given a corpus of D documents, where each document contains a set of N words, the word frequency can be represented by a D x N matrix, denoted by X, where $X_{d,n}$ is the frequency of word n in document d.

This vector can be used to identify the most important or relevant words in a text, which can be useful for text summarization, keyword extraction, and text classification. Additionally, it can also help to understand the general theme or topic of a text. High frequency words that occur in a specific context can be used to identify the topic of a text. Word frequency is often used in combination with other features such as word co-occurrence, part-of-speech tagging, and syntactic parsing to extract more meaningful information from a text.

- **Topic-word Relationship**

The topic-word relationship in topic modeling refers to the way in which each topic is associated with a specific set of words or terms. This relationship is established using algorithms such as Latent Dirichlet Allocation (LDA) which analyze the text in a corpus and assign probabilities to the different topics based on the words in the text. Topic-word relationship can be represented mathematically using a probability distribution, where the probability of a word belonging to a particular topic is proportional to the relevance of that word to the topic. This information can be used to identify the main themes or topics discussed in the corpus and how they are related to specific words or terms. This relationship is also used to understand the underlying structure of the text and how different topics are related to each other. It can also be used to identify the specific words that are most strongly associated with a topic, which can be useful for text summarization, keyword extraction, and text classification.

The mathematical notion of the topic-word relationship in topic modeling is typically represented by a matrix known as the "topic-word matrix" or "phi matrix". This matrix is created by applying a topic modeling algorithm such as Latent Dirichlet Allocation (LDA) to a corpus of documents. Each row of the matrix represents a topic, and each column represents a word or term in the vocabulary of the corpus. The entries of the matrix are the probabilities of a word belonging to a particular topic, which are calculated by analyzing the text in the corpus and determining the relevance of each word to each topic based on its occurrence in the text.

In more detail, the mathematical representation of topic-word relationship is defined as follows:

Given a corpus of D documents, where each document contains a set of N words, and K topics, the topic-word relationship is represented by a K x N matrix, denoted by φ, where φk,n is the probability of word n being generated by topic k. In other words, φk,n represents the probability of word n being in topic k. The entries in this matrix are non-negative and sum to one.

This matrix can be used to understand the underlying structure of the text and how different topics are related to specific words or terms. It can also be used to identify the specific words that are most strongly associated with a topic, which can be useful for text summarization, keyword extraction, and text classification.

In topic modeling, the joint likelihood is used to estimate the model parameters, such as the topic-word matrix and the document-topic matrix. The joint likelihood is the probability of observing the entire corpus of documents, given the model parameters. It is calculated by multiplying the individual probabilities of each word in each document, given the corresponding topic and the topic-word matrix. The mathematical notation for the joint likelihood can be represented as follows:

Given a corpus of D documents, where each document contains a set of N words, and K topics, the joint likelihood is represented as:

$$P(X \mid \theta, \varphi) = \prod_{d=1 \text{ to } D} \prod_{n=1 \text{ to } N} P(X_{d,n} \mid \theta_d, \varphi) \quad (1)$$

Where X is the word frequency matrix, θ is the document-topic matrix, φ is the topic-word matrix, Xd,n is the frequency of word n in document d, θd,k is the probability of topic k in document d and φk,n is the probability of word n being generated by topic k.

Estimating the model parameters, such as the topic-word matrix and the document-topic matrix, that maximize the joint likelihood, is commonly done by using optimization algorithms such as variational inference or Markov Chain Monte Carlo (MCMC). The joint likelihood is a key component for evaluating the goodness of the fit of the model, and comparing different models with different numbers of topics, or different parameterizations.

In the final step of LDA, each extracted topic is assigned a unique identifier or name for interpretation. The most reasonable terms in a topic are often analyzed in a ranked list of 3 to 30 terms to comprehend. However, this approach can present issues since it might be challenging to distinguish between the meanings of the themes because popular and non-descriptive phrases from the corpus frequently appear on the list. The term-topic relationship approach, developed by Sievert and Shirley [42], is employed to overcome this problem. By computing a weighted sum of the term's probability mathematics and the ratio of the term's probability inside a subject to its marginal probability across the corpus, this technique assesses a term's relevance to a topic. Instead of sorting phrases based on relevancy, this produces more cohesive and understandable subjects.

### 3.6 Language Analysis

Language analysis is a broad term that encompasses various techniques and approaches used to understand and interpret natural language texts or speech. It is a field that draws on various disciplines such as linguistics, computer science, and artificial intelligence.

The main goal of language analysis is to extract meaningful information from natural language texts and to understand the underlying structure and meaning of the text. To achieve this, various techniques are used, such as natural language processing, text mining, computational linguistics, and machine learning. Some of the specific tasks that language analysis can be used for include:

- Text classification: classifying text into predefined categories.
- Sentiment analysis: determining the sentiment or emotional tone of a text.
- Named entity recognition: identifying and extracting specific elements such as people, locations, and organizations from a text.
- Part-of-speech tagging: identifying the grammatical role of each word in a sentence.
- Parsing: analyzing the syntactic structure of a sentence
- Language modeling: generating coherent and grammatically correct text.

Language analysis can be applied to a wide range of applications, from natural language understanding and automated customer service to speech recognition and machine translation.

Measuring the orientation of natural language with regard to risk and optimism on the basis of rules can be done using a rule-based approach. This would involve defining a set of rules, based on specific keywords and phrases, that indicate a positive, negative or neutral sentiment towards risk and optimism. For example, a rule-based approach could include the following steps:

- Define a list of positive keywords and phrases associated with optimism, such as "hopeful," "optimistic," "positive outlook," etc.
- Define a list of negative keywords and phrases associated with pessimism, such as "doubtful," "pessimistic," "negative outlook," etc.
- Define a list of keywords and phrases associated with risk, such as "risk," "danger," "uncertainty," etc.
- For each text, use regular expressions or other string-matching techniques to count the number of occurrences of each positive, negative, and risk keywords and phrases.
- Assign a sentiment score to the text based on the number of occurrences of each keyword and phrase. For example, a positive sentiment towards optimism and low risk could be assigned a score of +1, a negative sentiment towards pessimism and high risk could be assigned a score of -1, and a neutral sentiment towards both could.

Determining a text's sentiment or feelings is assessing the direction of natural language. Opinion mining or sentiment analysis are terms used to describe this. Sentiment analysis is a branch of computational linguistics and natural language processing (NLP) that tries to recognize and retrieve qualitative information from text documents. There are several methods to measure the orientation of natural language, including:

- Rule-based methods: which involve the use of predefined lists of words, grammatical structures, and regular expressions to identify and classify the sentiment of a text.
- Machine learning-based methods: which involve the use of machine learning algorithms to train a model on a labeled dataset of texts, and then use the trained model to classify new texts.
- Hybrid methods: which combine the rule-based and machine learning-based methods to improve the accuracy and robustness of the sentiment analysis.
- Neural Network-based methods: which involve the use of neural networks, specifically Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) to analyze the text and determine the sentiment.

There are several uses for sentiment classification. The choice of approach will rely on the specific procedure and the financial budget. Each method has benefits and drawbacks of its own, such as social media monitoring, brand reputation management, and customer service automation.

Let Rd represent the calculated risk score and Od the optimism value for the document number d. We then calculate a ratio of the number of labeled words to the total number of words in the document, which is as follows.

$$R_d = \frac{\text{NoRW}}{TNoW} \tag{2}$$

$$O_d = \frac{\text{NoOW} - \text{NoPW}}{TNoW} \tag{3}$$

Where, NoRW is the number of risk words, TNoW is the total number of words, NoOW is the number of optimism words, and NoPW is the number of pessimisms words.

## 4 Empirical Setup and Settings

### 4.1 Study Setting

Our research specifically focuses on the energy sector, as it is currently undergoing significant changes and transformations in many countries. These changes include the reduction of greenhouse gas emissions and the emergence of new companies in the field, which have a significant impact on operations and strategies. For example, investment decisions in the energy sector are heavily influenced by factors such as risk and return and are also affected by public policies and regulations.

Financial filings serve as the foundation of our study since they offer a comparatively objective assessment of a company's current performance and the dangers that go along with it. These disclosures are especially helpful in recognizing the current issues and changes that firms in the energy industry are dealing with.

We collected US Securities & Exchange Form 8-K filings. Companies must submit these filings to update stakeholders on all current changes and events deemed important. Along with information particular to the energy industry, such as legal risks associated with policy changes, this also contains details about financial success, securities sales, buyouts, and organizational restructuring. Unforeseen occurrences, like the effects of the Deepwater tragedy in 2010, are also covered in Form 8-K filings. Form 10-K and Form 10-Q, which give monthly and quarterly income statements, differ from this filing. Form 8-K filings should also be scheduled so management can take quick strategic action. Additionally, the filings provide the industry in which the business works, allowing us to easily identify and select businesses that are pertinent to the electricity sector for our study.

### 4.2 Statistics Analysis

We examined a dataset that includes all publicly traded stocks on the NYSE from 2004 to 2021, consisting of 280,636 filings. We applied several filters to this dataset, including removing filings that couldn't be matched to stock symbols in order to obtain information about the stock market reaction of investors. We also excluded filings with less than 200 words, as per the methodology of Loughran and McDonald. These filters left us with a final corpus of 266,416 filings, of which 13% were from the energy sector.

In the corpus of 266,416 files, we examined the frequency and duration of the filings. The corpus consists of files from 180 businesses engaged in coal mining and offshore drilling, among other energy-related subsectors. The largest provider of oil field services worldwide, Schlumberger Ltd., is among its mid-sized businesses and international players. There are 136 filings on average per firm, with the STD of 92.19. A corporation may file as little as 1 document or as many as 482 documents. An individual file typically has 3961.82 words. Additionally, a modest trend in the frequency of submissions throughout each year was revealed by our analysis. Figure 2 shows the number of 8-K filings from 2004 to 2021. The colored bars show the number of submissions by firms in the energy industry, while the white bars show the overall number of submissions throughout the research period.

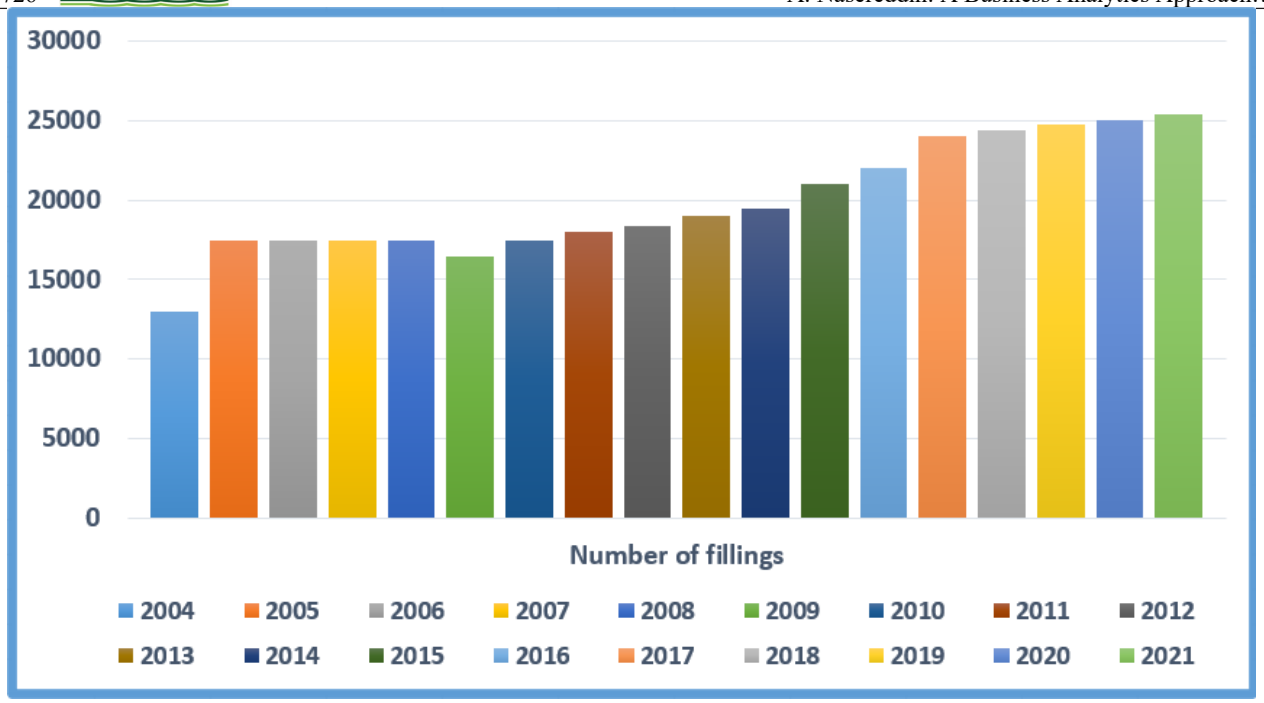


**Fig. 2.** shows the number of 8-K filings from 2004 to 2021. The colored bars show the number of filings by firms in the energy industry, while the white bars show the overall number of submissions throughout the research period.

## 5 Empirical Findings

In this part, we examine the available information. In order to determine the main topics in the story materials from firms in the energy industry, we first undertake topic modeling. We then look at how words that suggest dangers or optimism are used in relation to each topic that has been discovered. The basis for additional both internal and external assessments is laid forth here.

### 5.1 Identification of Topics

We performed the classification technique using Latent Dirichlet Allocation (LDA), but we had to decide how many topics to discover. In contrast to previous machine learning methods, this one does not use cross-validation or heuristics to optimize the number of groups. Five individuals from our company were given the top 10 most pertinent terms from subject models with various numbers of topics (k = 5, 10, 20, 30, 40, 50) in order to identify the ideal number of subjects. They were instructed to provide names to each subject and select the value of k that produced the most highly coherent particular subjects that were mutually exclusive yet collectively exhaustive. The students frequently concurred that the ideal number of subjects was 22, supported by related studies [43, 44]. Our managerial tool also allows one to alter the number of subjects and regulate the level of analytical granularity. As part of our sensitivity analysis, we conducted further empirical tests with different configurations and got positive findings that validated our hypotheses.

We employed a two-stage process once the tests were finished to give each retrieved theme a special name. First, we derived each subject's titles from the phrases that appeared most frequently in each topic. For instance, stemmed terms like "director," "appoint," "vote," and "elect" imply a subject pertaining to modifications in management or corporate governance. Likewise, word stems indicating financial outcomes disclosures include "quarter," "income," and "earnings". Second, we carefully reviewed sample files from our dataset to verify the subject names thoroughly, and we then asked our students to identify the topic names for the top 10 most pertinent terms. The five students had a high level of agreement as a consequence of this procedure, as indicated by the reasonably high inter-rater reliability, as expressed by Fleiss's kappa of 0.544. Table 1 lists the topic titles along with the number of fills. Table 2 displays the findings for each of the themes. Figure 3 for the danger and Figure 4 for the optimism show further explanations for these findings.

**Table 1:** The used topics names with the number of fillings.

| No. | Topic name | No. of filings |
|---|---|---|
| 1 | Loan arrangement topic | 5243 |
| 2 | Trust indenture topic | 2532 |

| | | |
|---|---|---|
| 3 | Legal issues topic | 112 |
| 4 | Earnings results topic | 1601 |
| 5 | Income statements topic | 151 |
| 6 | Security agreement topic | 74 |
| 7 | Employment agreement topic | 461 |
| 8 | Purchase agreement topic | 832 |
| 9 | Tax report topic | 351 |
| 10 | Stock option award topic | 339 |
| 11 | Resource development topic | 501 |
| 12 | Management change topic | 981 |
| 13 | Amendment of shareholder rights topic | 731 |
| 14 | Production outlook topic | 2981 |
| 15 | Infrastructure and logistics topic | 539 |
| 16 | Partnership arrangement topic | 734 |
| 17 | Mergers and acquisitions topic | 1422 |
| 18 | Public relations topic | 759 |
| 19 | Dividend payment topic | 1565 |
| 20 | Drilling contracts topic | 1089 |

**Table 2:** The results of the individual topics in terms of risk and optimism

| **No. of filings** | **Risk** | | | **Optimism** | | |
|---|---|---|---|---|---|---|
| | **Mean** | **Median** | **SD** | **Mean** | **Median** | **SD** |
| **1** | −0.32 | −0.34 | 1.00 | 0.34 | 0.37 | 0.91 |
| **2** | 0.72 | 0.72 | 0.70 | −0.11 | −0.08 | 0.62 |
| **3** | −1.13 | −0.96 | 0.44 | −0.35 | −0.53 | 0.69 |
| **4** | −0.81 | −0.84 | 0.85 | 0.62 | 0.74 | 0.76 |
| **5** | −0.82 | −0.84 | 0.84 | −0.34 | −0.54 | 0.95 |
| **6** | 0.11 | −0.06 | 0.95 | −1.13 | −0.92 | 1.04 |
| **7** | 0.41 | 0.42 | 0.81 | −0.36 | −0.28 | 1.19 |
| **8** | 0.52 | 0.66 | 0.83 | −1.36 | −1.45 | 1.51 |
| **9** | −0.23 | −0.30 | 0.61 | −1.85 | −1.91 | 0.85 |
| **10** | −0.82 | −0.97 | 1.22 | 0.18 | 0.09 | 0.81 |
| **11** | −0.54 | −0.65 | 0.71 | −1.34 | −1.24 | 0.96 |
| **12** | 0.68 | 0.66 | 0.88 | 0.61 | 0.62 | 0.89 |
| **13** | 0.22 | 0.32 | 0.64 | −0.05 | −0.16 | 0.62 |
| **14** | 0.74 | 0.64 | 0.81 | 0.08 | 0.16 | 0.91 |
| **15** | −0.60 | −0.64 | 0.81 | 0.06 | 0.15 | 0.86 |
| **16** | −0.62 | −0.85 | 0.72 | −0.16 | −0.27 | 0.58 |
| **17** | −0.78 | −0.83 | 0.63 | −0.13 | −0.21 | 0.94 |
| **18** | −0.67 | −0.71 | 0.39 | −0.94 | −1.04 | 0.53 |
| **19** | 0.45 | 0.37 | 0.85 | 0.34 | 0.35 | 0.7 |
| **20** | 0.34 | 0.37 | 0.82 | 0.11 | 0.02 | 0.68 |

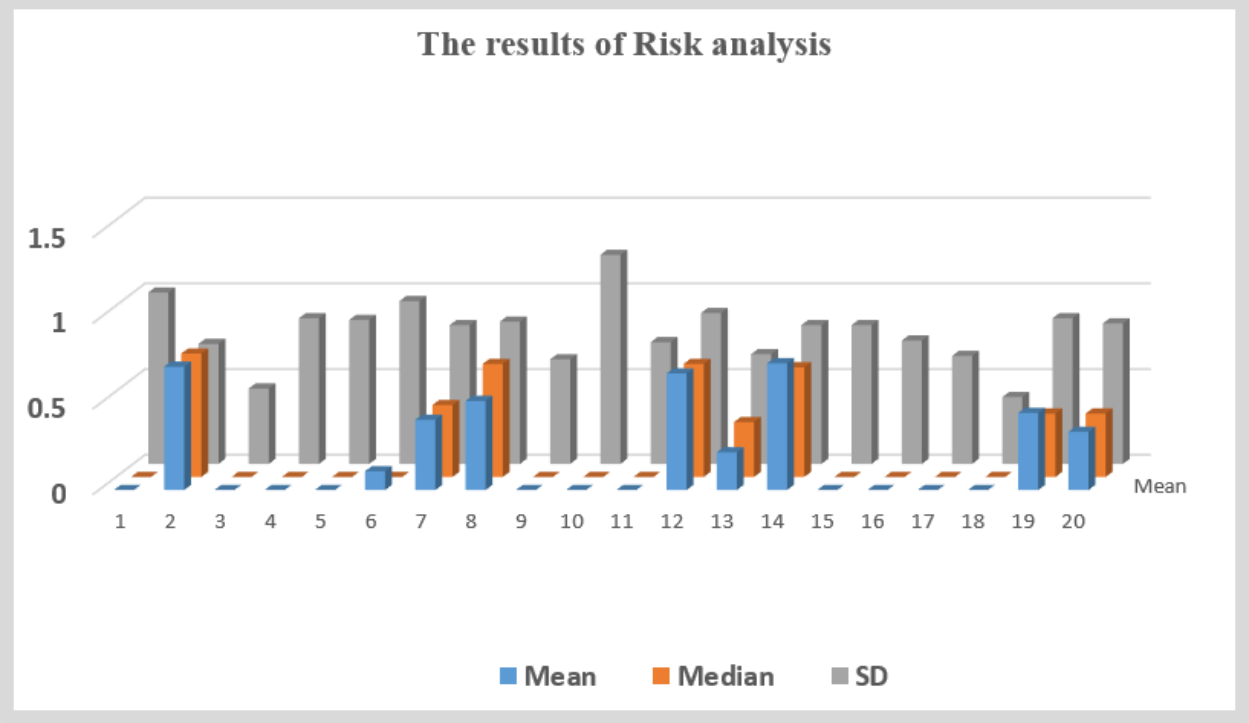


**Fig. 3.** The results of the individual topics in terms of risk.

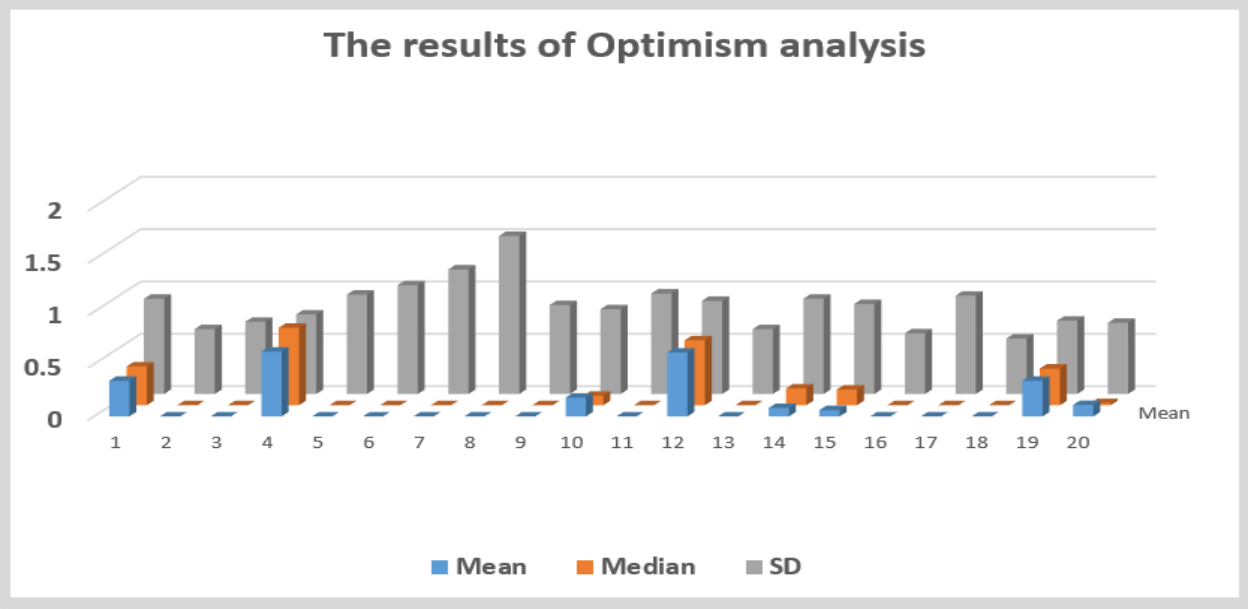


**Fig. 4.** The results of the individual topics in terms of Optimism

### *5.2 Descriptive statistics of topics*

The regularity and distribution of the various subjects are now more thoroughly examined. We allocated each file in our collection to the subject with the highest conditional distribution, as was already established. Each topic's frequency is shown in Table 2. Six topics—loan arrangement, trust indenture, profits results, production projection, acquisitions and acquisitions, and dividend payment—are assigned to most of the documents. Two-thirds of all files are comprised of these six themes, with the remaining topics being spread among the others. Financial documents frequently include profits outcomes, which is a high percentage. This conclusion is consistent with other studies [45].

The different issues are further divided into categories in Table 2 based on the strategic characteristics of risk and optimism. For instance, the mean risk score for 8-K filings related to a specific issue is displayed in the fourth column. As we can see, there are large differences in the identified topics' median and standard deviation. For instance, the mean

level of risk for the issue of production forecast is rather high, but the risk level for the topic of earnings performance is often lower. The optimism measure exhibits a similar trend. As we can see, each issue has a varied amount of volatility for both the risk and optimism scores. For instance, as seen in Figures 3 and 4, the issue of managerial change transmits a high degree of optimism. However, the average positivity level for purchase agreements is quite low. For instance, the optimism measure for a purchase agreement is 1.49, whereas it is lower for an investor protection modification.

### *5.3 Strategic analysis of the internal environment*

A strategic analysis of the internal environment involves evaluating a company's internal resources, capabilities, and competencies to identify strengths and weaknesses that can impact its ability to achieve its goals and objectives. This analysis can help a company identify opportunities for improvement and areas where it may need to invest in resources or develop new capabilities.

With the help of our technique, practitioners may plot key metrics on a risk-optimism matrix to produce a simple and clear visual depiction of the outcomes. As an illustration, we will demonstrate how we used this methodology to examine all filings made by two distinct corporations between the years 2004 and 2021.

Our approach quickly identifies variations in how various news subjects are communicated, enabling businesses to assess their strengths and shortcomings. The world's largest publicly listed worldwide oil and gas corporation, ExxonMobil, illustrates how our methodology was used. ExxonMobil is one of the country's largest manufacturers, marketers of petroleum products, and chemical producers in the world, and it also has a resource inventory that leads the industry. For instance, files pertaining to mergers and acquisitions have a negative tone and a medium risk rating. ExxonMobil, in contrast, uses a highly upbeat tone and a relatively low-risk level when releasing press announcements on dividend payments and profit outcomes. This is conceivable given that ExxonMobil's market capitalization more than quadrupled throughout our investigation. In addition, it ranked as the second-most lucrative business in the Fortune 500 in 2014.

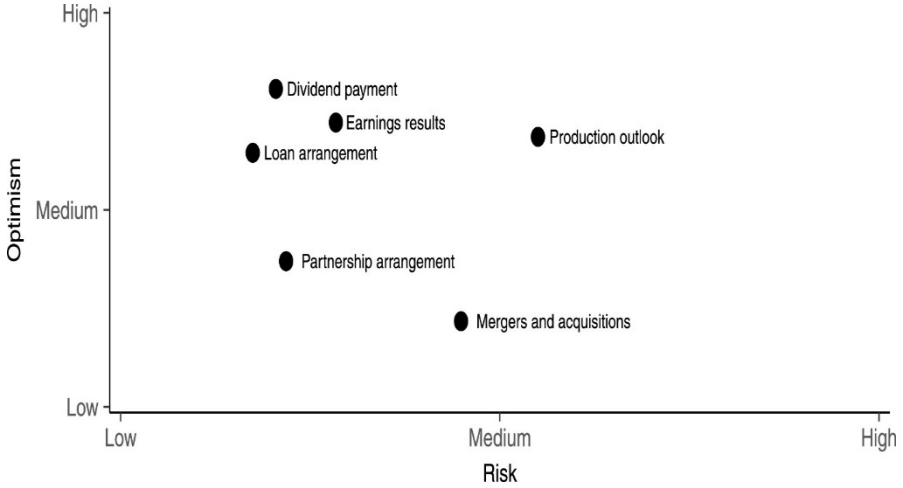


**Fig. 5.** Matrix displaying the level of risk and optimism in news filings from ExxonMobil.

We give a second risk-optimism matrix that examines the same themes for Transocean Ltd., one of the biggest offshore drilling contractors in the world, to contrast the analysis of ExxonMobil. This firm offers offshore agreement services worldwide for oil and gas wells worldwide and has operations in more than 20 nations. As seen in Figure 6 [7], compared to ExxonMobil, we instantly notice disparities in the transmission of several news subjects. For instance, the tone and risk level of filings relating to dividend payment and output projection is extremely gloomy. This is important since one of Transocean Ltd.'s oil rigs exploded in the Gulf of Mexico, causing the Deepwater Horizon oil disaster. Our results are consistent with the stock market valuation of Transocean Ltd., whose stock price declined from an all-time high of $160 in 2008 to a record-breaking low of $7.55 in 2021.

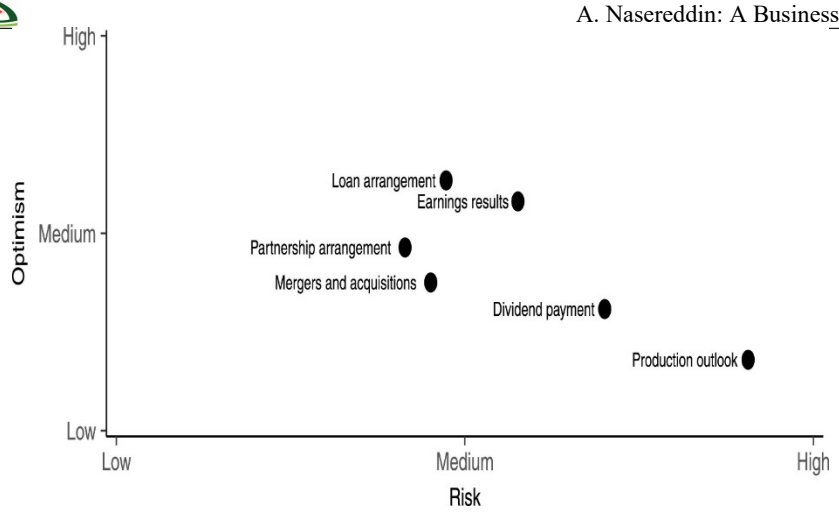


**Fig. 6.** Matrix displays the level of risk and optimism in news filings from Transocean Ltd.

Although our method eliminates many of the drawbacks of manual strategic frameworks, it has several limitations. Language may be ambiguous, and in finance, managers can feel compelled to construct their disclosures in a particular way. For instance, they could substitute positive remarks for negative ones, which might make it challenging to infer meaning correctly. Additionally, according to behavioral studies, text mining can only estimate writers' subjective opinions and cannot accurately transform linguistic data into numerical evaluations. Future study is needed in this area, in any case. Furthermore, our method can only detect the company's performance as it is contained in narrative materials. Therefore, it is important to take precautions to get beyond any restrictions imposed by the news source of choice. Combining several media sources, such as financial disclosures, press announcements, or corporate disclosures, which can provide internal data while newspaper headlines and user evaluations present an external perspective, is an option.

## 6 Conclusion and Future Works

In conclusion, business analytics for strategic management using topic modeling is a powerful tool for organizations to gain a better understanding of their internal and external environments. Topic modeling is a technique that can be used to extract insights from large sets of unstructured data, such as customer feedback, social media posts, and news articles. By applying topic modeling to these data sources, organizations can gain a better understanding of their customers, competitors, and market trends, which can be used to make strategic decisions such as identifying new business opportunities, improving product or service offerings, or identifying potential risks. Additionally, by combining topic modeling with other business analytics techniques such as predictive modeling, organizations can gain a more comprehensive view of their performance and make data-driven decisions. Overall, topic modeling is a valuable tool for strategic management and can provide organizations with insights to help them stay ahead of the competition and make better decisions.

SWOT analysis and other management methods are frequently used in strategic planning. Even though it has been around for a while, SWOT analysis is still a useful tool, and any attempt to simplify its study may benefit businesses and organizations directly. Our method uses the most current advancements in text mining and technical analysis, especially by inferring problems with the order for businesses from narrative materials and varying the risk-strength scores given to these structures depending on the syntax. This serves as an early warning system for significant developments and enables tracking key performance indicators in key areas. Additionally, it enables external assessments of rivals and the broader market environment. Automation, repeatable calculation methods, and a variety of granularity degrees, from sector units to specific processes and activities, are advantages of this approach. Our evaluations may assist managers in creating and adjusting their plans according to the results, making it a very adaptable strategy that can be used for random businesses, industries, and disciplines.

There are several future work directions for Business analytics for strategic management using topic modeling:

- Combining different data sources: Combining different data sources such as social media, customer feedback, and news articles, in addition to financial disclosures, can provide a more comprehensive understanding of a company's internal and external environments.
- Incorporating other NLP techniques: Incorporating other natural language processing techniques such as sentiment analysis and named entity recognition can provide additional insights into the language used in narrative materials.

- Automating the topic labeling process: Developing an automated method for labeling topics can reduce the time and effort required for manual labeling.
- Incorporating other business analytics techniques: Combining topic modeling with other business analytics techniques such as predictive modeling can provide a more comprehensive understanding of a company's performance and make data-driven decisions.
- Conducting case studies and experiments: Conducting case studies and experiments using real-world data can help to further validate and refine the proposed framework.
- Increasing the granularity of the analysis: Increasing the granularity of the analysis to identify key issues at the level of individual business units, activities and processes.
- Developing a management tool: Developing a management tool that can be used by organizations to apply the proposed framework to their own data.

**Ethical approval:**

This article does not contain any studies with human participants or animals performed by any of the authors.

**Informed consent:**

Informed consent was obtained from all individual participants included in the study.

**Data availability statements:**

Data is available from the authors upon reasonable request.

**Funding:**

Not Applicable

**Conflict of interest:**

The authors declare that there is no conflict regarding the publication of this paper.

## References

[1] Gandomi, A.H., F. Chen, and L. Abualigah, Machine learning technologies for big data analytics. 2022, MDPI. p. 421.

[2] Al-Sai, Z.A., et al., Explore Big Data Analytics Applications and Opportunities: A Review. Big Data and Cognitive Computing, 2022. 6(4): p. 157.

[3] Daoud, M.S., et al., Gradient-Based Optimizer (GBO): A Review, Theory, Variants, and Applications. Archives of Computational Methods in Engineering, 2022: p. 1-19.

[4] Wu, D., et al., Modified Sand Cat Swarm Optimization Algorithm for Solving Constrained Engineering Optimization Problems. Mathematics, 2022. 10(22): p. 4350.

[5] Elbashir, M.Z., P.A. Collier, and M.J. Davern, Measuring the effects of business intelligence systems: The relationship between business process and organizational performance. International journal of accounting information systems, 2008. 9(3): p. 135-153.

[6] Vidgen, R., S. Shaw, and D.B. Grant, Management challenges in creating value from business analytics. European Journal of Operational Research, 2017. 261(2): p. 626-639.

[7] Pröllochs, N. and S. Feuerriegel, Business analytics for strategic management: Identifying and assessing corporate challenges via topic modeling. Information & Management, 2020. 57(1): p. 103070.

[8] Helms, M.M. and J. Nixon, Exploring SWOT analysis–where are we now? A review of academic research from the last decade. Journal of strategy and management, 2010.

[9] Dess, G.G. and G.T. Lumpkin, The role of entrepreneurial orientation in stimulating effective corporate entrepreneurship. Academy of Management Perspectives, 2005. 19(1): p. 147-156.

[10] Waggoner, D.B., A.D. Neely, and M.P. Kennerley, The forces that shape organisational performance measurement systems:: An interdisciplinary review. International journal of production economics, 1999. 60: p. 53-60.

[11] Kaplan, R.S. and D.P. Norton, Transforming the balanced scorecard from performance measurement to strategic management: Part II. Accounting horizons, 2001. 15(2): p. 147-160.

[12] Hatten, M.L., Strategic management in not-for-profit organizations. Strategic Management Journal, 1982. 3(2): p. 89-104.

[13] Hannigan, T.R., et al., Topic modeling in management research: Rendering new theory from textual data. Academy of Management Annals, 2019. 13(2): p. 586-632.

[14] Kim, J. and Y. Geum, How to develop data-driven technology roadmaps: The integration of topic modeling and link prediction. Technological Forecasting and Social Change, 2021. 171: p. 120972.

[15] Jeyaraj, A. and A.H. Zadeh, Evolution of information systems research: Insights from topic modeling. Information & Management, 2020. 57(4): p. 103207.

[16] Mustak, M., et al., Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda. Journal of Business Research, 2021. 124: p. 389-404.

[17] Bodendorf, F., B. Wytopil, and J. Franke, Business Analytics in Strategic Purchasing: Identifying and Evaluating Similarities in Supplier Documents. Applied Artificial Intelligence, 2021. 35(12): p. 857-875.

[18] Fan, S., R.Y. Lau, and J.L. Zhao, Demystifying big data analytics for business intelligence through the lens of marketing mix. Big Data Research, 2015. 2(1): p. 28-32.

[19] Yang, H.-L., T.-W. Chang, and Y. Choi, Exploring the research trend of smart factory with topic modeling. Sustainability, 2018. 10(8): p. 2779.

[20] Jelodar, H., et al., Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 2019. 78: p. 15169-15211.

[21] Bromiley, P. and D. Rau, Towards a practice-based view of strategy. Strategic Management Journal, 2014. 35(8): p. 1249-1256.

[22] Capon, N., J.U. Farley, and J.M. Hulbert, Strategic planning and financial performance: more evidence. Journal of management studies, 1994. 31(1): p. 105-110.

[23] Vaara, E. and J.-A. Lamberg, Taking historical embeddedness seriously: Three historical approaches to advance strategy process and practice research. Academy of Management Review, 2016. 41(4): p. 633-657.

[24] Alba, J.W. and J.W. Hutchinson, Dimensions of consumer expertise. Journal of consumer research, 1987. 13(4): p. 411-454.

[25] Jarzabkowski, P. and S. Kaplan, Strategy tools-in-use: A framework for understanding "technologies of rationality" in practice. Strategic management journal, 2015. 36(4): p. 537-558.

[26] Wooldridge, B. and S.W. Floyd, Research notes and communications strategic process effects on consensus. Strategic Management Journal, 1989. 10(3): p. 295-302.

[27] O'Reilly, C.A., et al., How leadership matters: The effects of leaders' alignment on strategy implementation. The leadership quarterly, 2010. 21(1): p. 104-113.

[28] Powell, T.C., D. Lovallo, and C.R. Fox, Behavioral strategy. Strategic Management Journal, 2011. 32(13): p. 1369-1386.

[29] Dess, G.G., Strategic management: Text and cases. 2007: Mc Graw Hill.

[30] Bishop, C.M. and N.M. Nasrabadi, Pattern recognition and machine learning. Vol. 4. 2006: Springer.

[31] Tetlock, P.C., M. Saar-Tsechansky, and S. Macskassy, More than words: Quantifying language to measure firms' fundamentals. The journal of finance, 2008. 63(3): p. 1437-1467.

[32] Loughran, T. and B. McDonald, Textual analysis in accounting and finance: A survey. Journal of Accounting Research, 2016. 54(4): p. 1187-1230.

[33] Hanley, K.W. and G. Hoberg, Litigation risk, strategic disclosure and the underpricing of initial public offerings. Journal of Financial Economics, 2012. 103(2): p. 235-254.

[34] Ravi, K. and V. Ravi, A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge-based systems, 2015. 89: p. 14-46.

[35] Nusir, M., et al., Design Research Insights on Text Mining Analysis: Establishing the Most Used and Trends in Keywords of Design Research Journals. Electronics, 2022. 11(23): p. 3930.

[36] Abualigah, L. and K.H. Almotairi, Dynamic evolutionary data and text document clustering approach using improved Aquila optimizer based arithmetic optimization algorithm and differential evolution. Neural Computing and Applications, 2022. 34(23): p. 20939-20971.

[37] Ikotun, A.M., et al., K-means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data. Information Sciences, 2022.

[38] Ezugwu, A.E., et al., A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Engineering Applications of Artificial Intelligence, 2022. 110: p. 104743.

[39] Loughran, T. and B. McDonald, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of finance, 2011. 66(1): p. 35-65.

[40] Manning, C. and H. Schutze, Foundations of statistical natural language processing. 1999: MIT press.

[41] Blei, D.M., Probabilistic topic models. Communications of the ACM, 2012. 55(4): p. 77-84.

[42] Sievert, C. and K. Shirley. LDAvis: A method for visualizing and interpreting topics. in Proceedings of the workshop on interactive language learning, visualization, and interfaces. 2014.

[43] Ramage, D., et al. 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. 2009.

[44] Niederhoffer, V., The analysis of world events and stock prices. The Journal of Business, 1971. 44(2): p. 193-219.

[45] Carter, M.E. and B.S. Soo, The relevance of Form 8-K reports. Journal of Accounting Research, 1999. 37(1): p. 119-132.