Essi Nousiainen

Appearance of Corporate Innovation in Financial Reports

A Text-Based Analysis

School of Accounting and Finance
Master's Thesis in Economics
Master's Degree Programme in Economics

Vaasa 2020

University of Vaasa
School of Accounting and Finance
Author: Essi Nousiainen
Title of the Thesis: Appearance of Corporate Innovation in Financial Reports : A Text-Based Analysis
Degree: Master of Science in Economics and Business Administration
Programme: Master's Degree Programme in Economics
Supervisor: Jaana Rahko
Year of Graduation: 2020 **Pages:** 76

ABSTRACT:
Innovations are important drivers of economic growth and firm profitability. Firms need funding to generate profitable innovations, which is why it is important to reliably distinguish innovative firms. Innovation indicators are used to measure this innovativeness, and consequently, it is important that the used indicator is reliable and measures innovation as desired.

Patents, research and development expenditure and innovation surveys are examples of popular innovation indicators in research literature. However, these indicators have weaknesses, which is why new innovation indicators have been developed. This thesis studies the text-based innovation indicator developed by Bellstam et al. (2019) with a new type of data. Bellstam et al. (2019) created a new text-based innovation indicator that compares corporations' analyst reports with an innovation textbook as the basis for the indicator. The similarity between these texts created the measurement for innovativeness. Analyst reports are usually subject to charge. However, the 10-K reports used as data for this study are publicly available, and their functionality as the basis of the innovation indicator would mean good availability for the indicator.

The study begins by training a Latent Dirichlet allocation (LDA) model with a sample of 10-K documents from 2008-2018. LDA-model is an unsupervised machine learning method, it finds topics in the text documents based on the probabilities of different words. The LDA-model was trained to find 15 topic allocations in the data and the output of the model is the distribution of these topics for each document. The same topic distributions were also allocated for eight samples from innovation textbooks. When the topic distributions were allocated, a Kullback-Leibler-divergence (KL-divergence) was calculated between each text sample and 10-K document. Thus, the KL-divergence calculated is the lowest for those reports that are the most similar to the innovation text and works as the text-based innovation indicator.

Finally, the text-based innovation indicator was validated with regression analysis, in other words, it was confirmed that the indicator measures innovation. The text-based indicator was compared with research and development costs and the balance sheet value of brands and patents in different linear regressions. Out of the eight innovation measurements, most had a statistically significant correlation with one or both of the other innovation indicators. The ability of the text-based indicator to predict the development of sales in the next year was studied with regression analysis as well and all of the measurements had a significant effect on this. The most significant findings of this thesis are the relationship of the text-based innovation indicator and other indicators and its ability to predict firms' sales.

KEYWORDS: Innovation, textual analysis, annual reports, economics, machine learning

VAASAN YLIOPISTO
Laskentatoimen ja rahoituksen yksikkö
Tekijä: Essi Nousiainen
Tutkielman nimi: Appearance of Corporate Innovation in Financial Reports : A Text-Based Analysis
Tutkinto: Kauppatieteiden maisteri
Oppiaine: Taloustiede
Työn ohjaaja: Jaana Rahko
Valmistumisvuosi: 2020 **Sivumäärä:** 76

TIIVISTELMÄ:
Innovaatiot ovat tärkeitä talouskasvun ja yritysten kannattavuuden ajureita. Tuottavien innovaatioiden syntymiseksi yritykset tarvitsevat rahoitusta, minkä takia onkin tärkeää, että innovatiiviset yritykset pystytään tunnistamaan luotettavasti. Innovaatioindikaattoreita käytetään tähän innovatiivisuuden mittaamiseen ja on siksi tärkeää, että käytetty indikaattori on luotettava ja mittaa innovatiivisuutta oikealla tavalla.

Kirjallisuudessa paljon käytettyjä innovaatioindikaattoreita ovat esimerkiksi patentit, tutkimus- ja kehitysmenot sekä innovaatiokyselyt. Näissä indikaattoreissa on kuitenkin myös heikkouksia, joiden takia uusia indikaattoreita on alettu kehittää. Tässä tutkielmassa tutkitaan Bellstamin ja muiden (2019) luomaa tekstipohjaista innovaatioindikaattoria erilaisella datalla. Bellstam ja muut (2019) loivat uuden innovaatioindikaattorin, jonka pohjana oli yritysten analyytikkoraporttien vertailu innovaatio-oppikirjan tekstin kanssa, näiden samankaltaisuusvertailusta saatiin innovaatiomittari. Analyytikkoraportit ovat usein maksullisia. Tässä tutkimuksessa aineistona on käytetty lakisääteisiä tilinpäätösraportteja, jotka ovat julkisia tiedostoja, joten niiden toimivuus innovaatioindikaattorin pohjana tarkoittaisi hyvää saatavuutta indikaattorille.

Tutkimus alkaa Latent Dirichlet allocation (LDA) –mallin harjoittamisella Yhdysvaltalaisten yritysten 10-K, eli tilinpäätösraporteilla vuosilta 2008-2018. LDA-malli on valvomaton koneoppimismenetelmä, eli se etsii datasta itse aihepiirejä sanojen todennäköisyyksien perusteella. LDA-malli asetettiin etsimään datasta 15 eri aihepiiriä raporteissa käytettyjen aiheiden perusteella ja mallin tuloksena on näiden aihepiirien jakautuminen jokaisessa dokumentissa. Samat aihepiirijakaumat haettiin myös kahdeksalle tekstiotokselle innovaatio-oppikirjoista. Aihepiirijakaumien ollessa valmiit, laskettiin Kullback-Leibler-divergenssi (KL-divergenssi) tilinpäätösraporttien ja innovaatio-oppikirjojen tekstiotosten aihepiirijakaumien välille. Laskettu KL-divergenssi on siten matalin niille tilinpäätösraporteille, joiden teksti on lähimpänä kunkin innovaatio-oppikirjan tekstiä ja toimii tekstipohjaisena innovaatioindikaattorina.

Lopuksi indikaattorin toimivuus vahvistetaan regressioanalyysillä, eli tutkitaan, että se mittaa innovatiivisuuta. Regressioanalyysillä tutkitaan innovaatiomittarien yhteyttä yritysten tutkimus- ja kehitystoiminnan kuluihin sekä patenttien ja brändien tasearvoon. Kahdeksasta innovaatiomittarista suurimmalla osalla oli tilastollisesti merkitsevä yhteys muuttujista toiseen tai molempiin. Myös uuden innovaatiomittarin kykyä ennustaa yritysten seuraavan vuoden myyntiä tutkittiin regressioanalyysillä ja jokaisella mittarilla oli tilastollisesti merkitsevä yhteys yritysten liikevaihdon muutokseen. Tutkimuksen merkittävin löydös oli tekstipohjaisen innovaatiomittarin yhteys muihin innovaatiomittareihin ja yritysten liikevaihdon kehitykseen.

AVAINSANAT: Innovaatiot, kansantaloustiede, koneoppiminen, tilinpäätös, tekstianalyysi

Table of Contents

Figures

Tables

1 Introduction

Innovations are important in both, macro- and microeconomics, due to their effects on economic growth and firm profits. Innovative firms can generate a higher profit, which in turn increases total economic growth. Innovations also increase total factor productivity, due to more efficient production methods and positive externalities. It is important that innovative firms and projects get funding, which is why we need to be able to distinguish innovative firms from non-innovative firms. Innovation indicators are needed to reliably measure this distinction.

In current literature, many different types of proxies are used to measure innovation, the proxies include surveys, measures related to the inputs of innovation (e.g. research and development, R&D) and measures related to the outputs of innovation (e.g. patents) (Greenhalgh & Rogers, 2010, p. 58-62). All of the innovation indicators have their weaknesses related to the range of innovative activities that they are able to capture. The shortcomings of innovation indicators currently used in literature call for a more comprehensive way of measuring innovation.

The objective of this thesis is to study whether the innovativeness of a firm can be measured from the narrative sections of 10-K filings. The secondary objective is to form an innovation indicator based on textual analysis of these 10-K reports and test whether it correlates with innovation. A text-based innovation indicator has been studied successfully in previous literature, with analyst reports as the source text for the measurement (Bellstam et al., 2019).

New measures of innovation have been developed in the recent years to compete with the traditional indicators. In addition to the one developed by Bellstam et al. (2019), Mukherjee et al. (2017) introduced a text-based innovation measurement, where they studied the market response to innovation-related press releases. The innovation measure by Kogan et al. (2017), on the other hand, combined the stock market response to news about patents with patent data. Cooper et al. (2020) introduced an innovation

measure that is based on the output elasticity of R&D. This study aims to extend the literature on new innovation measurements.

The benefit of an innovation indicator based on financial report text would be that it could be measured for any firm, only if a financial report is available. Since it is mandatory for public companies to publish these reports, the availability of data is good for public companies. Making the measurement for private companies could be difficult though, since their reporting is usually mainly numeric.

The study is conducted by comparing the topic distributions extracted from the 10-K filings with topic distributions of text samples from innovation textbooks. The measurement of innovativeness is based on the similarity of these topic distributions. The method is then validated by comparing it with traditional innovation indicators and the growth of future income to establish, whether the measure that has been created, (1), captures innovation, and (2), is at least as good as the traditional innovation indicators.

Chapter 2 presents the theoretical basis for this study followed by the hypothesis. Chapter 3 deals with the theory and methods of text analysis and natural language processing (NLP). Chapter 4 presents the data and descriptive statistics, and possible issues with the selected data source. Research design and methods are elaborated in chapter 5. Research results can be read from chapter 6, and conclusions are found in chapter 7.

2 Theory and Research Hypothesis

2.1 Innovation Economics

2.1.1 Definition of Innovation

OECD (2005) defines innovation as such:

> *An innovation is the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organisational method in business practices, workplace organisation or external relations.*

Taques et al. (2020) write that in both, manufacturing and service industries, innovation can be a product, a process, a marketing or an organizational innovation. Product innovation can be the creation of a new product or service or an improvement of an existing one. Process innovations are improvements or alterations in production or delivery methods or service production. Marketing innovation can be, for example, new product design, and organizational innovations can be new business practices or new ways of physical composition of the company etc.

An innovation requires the element of novelty; Greenhalgh and Rogers (2010, p. 5) define it as new to the firm and to the relevant market, but they call innovation that is only new to the firm, imitation. Gordon and McCann (2005), however, leave the identification of innovation for the firm itself, because then the definition can be applied to different industrial sectors and product and process innovations equally. According to Atkinson and Ezell (2012, p. 129), novelty alone does not establish innovation though, since all inventions are not innovations, but innovation requires business application. Compared to inventions, innovations have been commercialized. Lastly, an innovation needs to be an improvement to the existing options, only broadening variation does not constitute innovation (Gordon & McCann, 2005).

According to Greenhalgh and Rogers (2010, p. 5), innovation can vary from incremental to drastic, incremental innovation is a small change in an existing product and drastic

innovation is a completely new method of production with a new genre of innovative products, such as the steam engine.

In conclusion, there are many different ways to be innovative and novelty is the common factor. Finding a measurement that can capture all different types of innovation can prove to be tricky. As discussed in the next chapter, literature has used different ways to measure innovation. The more traditional ways of measuring innovation have a common flaw taking into account only a fraction of innovation, which is why new, better indicators are necessary.

2.1.2 Macroeconomic Effects of Innovation

The role of innovation in economic growth has been a point of interest for a long time. Schumpeter (1943) discussed creative destruction caused by technological innovation in his book Capitalism, Socialism and Democracy. Kuznets (1969, ch. 1) discussed the role of innovations and exploitation of new knowledge in economic development throughout history. Aghion and Howitt (1998) later developed the theory of Schumpeterian growth based on Schumpeter's creative destruction.

Aghion and Howitt (1998, p. 11) argue that technological progress is necessary for long-run economic growth due to diminishing returns to capital. For example, giving a worker a hammer will increase his nailing productivity, but if the person gets ten more pieces of the same hammer, his productivity will not grow tenfold. This is a simple explanation as to why technological development is necessary for increasing productivity. To increase the productivity of the worker with a hammer, the worker needs a more efficient hammer (or a nail gun).

According to Howitt (2004), in endogenous growth theory, the determinant of long-run economic growth is total-factor productivity, which mainly depends on technological

progress. Technological progress for its part comes from innovation. The two types of endogenous growth theories are AK theory and Schumpeterian theory.

In their book, Aghion and Howitt (1998, p. 11-16) write that earlier growth theories, such as the Solow-Swan model, considered technological change as an exogenous variable to the model. These exogenous growth theories also recognized technological progress as the driver of long-term economic growth, but the original models only recognized it as an exogenous factor. Endogenous AK models considered technological progress as a form of capital accumulation, as in knowledge accumulating over time (Howitt, 2004).

According to Aghion and Howitt (1998, p. 53), growth stems from vertical innovations that result from research activities, in the Schumpeterian approach to economic growth. Creative destruction is a key term in this type of economic growth, it means that new innovations make old technology obsolete. Incumbent producers give way to new and more efficient ones, which is called the business-stealing effect. In Schumpeterian growth, innovation has both negative and positive externalities. There is a negative externality for inefficient producers, but a positive externality to future research. The Schumpeterian approach also assumes that all innovations are drastic and do not face competition from the previous generation of innovations.

2.1.3 Firm Level Effects of Innovation

Firms can benefit from innovativeness in different ways and successful innovations can enhance individual firms' position in the market. Process innovation can give a firm competitive advantage, if immaterial property rights exist (Greenhalgh & Rogers, 2010, p. 11). The process innovation will give the firm the opportunity to undercut the competitors and capture the market or licence the process innovation to other producers and collect royalties. In this scenario, firms have significant incentives to innovate due to increased profits following successful process innovations. According to Weiss (2003), firms engage

in process innovation when they have less competition to decrease costs, because they can act as a monopolist and a product innovation will not increase their profits.

Greenhalgh and Rogers (2010, p. 12-14) discuss the effects of product innovation at microeconomic level. If a firm makes a product innovation it can protect with a patent, it can escape competition and act as a monopolist to maximize profits. However, if the product innovation is only incremental and either creates a new variety or improves quality, a monopoly situation might not be formed. In this case, the firm could face a new, steeper slope of the demand curve with lower price elasticity. Hombert and Matray (2018) discovered that U.S. firms engaging in innovation activities were able to escape the competition and were less impacted by Chinese imports than their non-innovative peers.

Junge et al. (2016) found empirical support to the hypothesis that marketing innovation, together with product innovation, increases firms' productivity growth. They also discovered that neither of them alone increase productivity, which implies complementarity. New products need to be marketed in an innovative manner to gain success. Due to the benefits of marketing innovation to firms, including it in innovation measurement would be justified. Marketing innovation cannot be measured through patents or other traditional indicators very easily, and thus, an indicator that captures a broader range of innovation would be necessary.

Camilsón and Villar-López (2014) studied the effect of organizational innovation on technological innovation. They found that organizational innovation is beneficial for technological innovation and both lead to an improvement in firm performance. The need for ways of including organizational innovation in innovation measurement is justified for the same reasons as for marketing innovation.

According to Aghion et al. (2018), the escape-competition effect in the Schumpeterian growth theory affects sectors where firms compete at the same technological level. In

these sectors, competition reduces surplus before innovation, and consequently, innovation leads to higher incremental profits and firms have the incentive to strive for the position of the market leader. In industrial sectors, where the technological level of the firms is uneven, there is a Schumpeterian effect, which decreases the incentives to innovate because the laggard firms' surplus decreases post-innovation. Hoberg and Phillips (2016) found supporting evidence that engaging in R&D activities increases product differentiation and firm profitability.

2.2 Indicators of Innovation

This chapter will take a closer look at ways to measure innovation. Mainly patents and research and development will be discussed due to their popularity in research, but many other indicators are at use too. Most innovation indicators are only able to capture a specific fraction of innovation on their own and due to this, innovation research nowadays focuses mainly on new product innovations (Bellstam et al., 2019). However, in the pursuits of capturing a wider range of innovation, composite indicators that summarize the information of various different indicators for a better overall view are also at use in research (Belitz et al., 2011).

Dziallas and Blind (2019), identified 82 different indicators of innovation from literature in the years between 1980 and 2015. Some indicators mentioned are patents and patent applications, research and development related indicators, the number of ideas, the ideas with commercialization potential, customer orientation, the number of new products and the success rate of new products. Examples of studies that either measure innovation using patents or evaluate the usefulness of patents as a proxy for innovation are Guan & Chen (2010), Bayarcelik & Tasel (2012), Belenzon & Patacconi (2013), Roper and Hewitt-Dundas (2015) and Dang & Motohashi (2015). Studies focusing on research and development input as an indicator of innovation include Belitz et al. (2011), Chiesa et al. (2009) and Edison et al. (2013).

2.2.1 Patents

Firms can use patents to obtain a temporary monopoly in the use of an invention (Belenzon & Patacconi, 2013). A patent can therefore strengthen the position of its owner in the market via more bargaining power, exclusivity or licensing income. A large patent portfolio can also increase firm value. However, patenting is quite expensive even though it can lead to the mentioned monetary benefits. Hall et al. (2005) found a positive relationship between the patent citations and market value of the firm, indicating that patents are focal elements of the intangible assets.

The World Trade Organization TRIPS agreement aims to ensure similar patent protection in all member countries (Hall & Harhoff, 2012). The objective of the TRIPS is to secure at least minimal patent protection and that product and process innovations regardless of the field of technology can gain patent protection for at least 20 years.

The limitations of patentability still slightly differ by country. In Finland, an invention that is new, inventive and has industrial application can be patented (PRH, 2019). Not everything can be patented though; according to the Finnish patent and registration office, discoveries, scientific theories, mathematical methods, aesthetic creations, schemes and methods for playing games or doing business, programs for computers and treatment methods practised on humans or animals cannot be patented. Inventions related to these can, however, be patented but only if they are technological in nature. Since patentability has limitations, patent data might not be fully trustworthy for measuring innovation. The patent law of the United States is slightly different; it requires usefulness, novelty and non-obviousness (USPTO, 2015). There are three types of patents in the United States, utility patents for a process, a machine, an article of manufacture or the composition of matter, design patents and plant patents.

Patents can be used as an indicator of innovative activity in firms and patent data has good availability (Griliches, 1990). Since Griliches' study, patents have been a popular

and common way to measure innovation in economic research. The weakness of patents as indicators of innovation is that not all inventions are patented and not all patents are of the same value (Nagaoka et al., 2010). Hall et al. (2013) found that out of all registered firms in the UK, only 1.6% used the patent system and out of those that engage in research and development only 4% applied for patents during 1998-2006. Out of different types of innovation for example organizational innovations cannot be patented. According to the study by the European Patent Office and the European Union Intellectual Property Office (2019), out of the 20 most patent-intensive industries, 17 are manufacturing industries measured by the amount of patents per 1,000 employees. The manufacturing industry generally uses patents, but many service-related innovations cannot be patented and therefore service-industry related innovation could be better measured by other means.

The WIPO (2020) International Patent Classification is a model of universal patent classification established to provide a search tool to efficiently find patent documents. The Classification standardizes patent documentation and ensures patent data availability, which could explain the popularity of patents as an innovation proxy. The patent document holds a lot of information about the patent. According to Nagaoka et al. (2010), the patent document's structure is the following: "the bibliographic information, the abstract of the information, the claims, the description of the invention, and the drawings and their description." The patent document also identifies the inventor and the applicator of the patent. The IPC ensures the availability of patent information and is the largest database with the broadest range of patent information.

2.2.2 Research and Development

Research and development expenditure is an indirect innovation measurement, since it only measures the input on innovative activities (Hong et al., 2012). Engaging in research and development activities can increase firms' innovative capacity through learning by

doing (Zhu et al., 2019). Research and development budget can be used to evaluate innovativeness with the assumption that firms with a higher R&D budget are more innovative (Dziallas & Blind, 2019).

As Greenhalgh and Rogers (2010, p. 59) say, R&D expenditure is a common indicator of innovation in research. Research and development are the inputs needed to produce innovation and patents, which is why they are used as an innovation proxy. However, Using R&D to predict innovation is not that straightforward. One of the issues of R&D expenditure is that it cannot predict the time of innovation and a time lag is possible.

The inherent uncertainty of R&D makes its innovation predicting ability questionable, R&D inputs on their own do not give a good estimate for firm innovativeness due to the uncertainty of them leading to a successful innovation (Cohen et al., 2013). R&D can lead to “good” and valuable innovation, but it can also lead to “bad” innovation. An innovation indicator which takes bad innovations into account as innovativeness might not be very useful for research purposes. In addition, unlike patents, R&D-data availability varies and is generally poorer for private firms (Cooper et al., 2020).

2.2.3 Surveys

According to Hong et al. (2012), innovation surveys are the commonly accepted innovation measure of today. Innovation is a spectrum of activities, which the surveys attempt to capture better than proxy measures like patents and R&D. Especially process- and organizational innovation, which the surveys are able to measure, are poorly represented by patent and R&D data.

One broad and ambitious survey is the EU Community Innovation Survey (CIS). The EU member states conduct CIS’s to gather innovation data (European Commission, 2020). The CIS is harmonized and voluntary to the member states and the surveys are carried

out every two years. Its objective is to provide information on the different types of innovation and the development of innovations. Other countries, such as the United States, Canada, Malaysia, Taiwan and South Korea, are utilizing or developing innovation surveys as well (Hong et al., 2012).

Innovation surveys come with their own issues, they are prone to human error and bias, and the representativeness of the survey depends on the response rate (Hong et al., 2012). The latest CIS survey period was 2016-2018 (Statistics Finland). The survey is a yes-no questionnaire, even though it would be more informative to measure the amount and quality of patents and R&D activities. The survey also has questions about product, service and process innovations, for product and service innovations, new-to-the-firm and new-to-the-market innovations are distinguished, but this division is not made for process innovations. Based on this, some improvements could be made to make the survey more informative, but on the other hand, this could raise the threshold to answer, which is probably why the questions are being kept simple. The survey does give good information about innovation at the firm level though, since it is answered by firm representatives.

2.2.4 Text-Based Approach

New methods of measuring innovation have been attempted to develop, since there are issues regarding the currently popular methods. Among them are text-based methods, which employ natural language processing (NLP) or text mining techniques to measure innovativeness, e.g. from analyst reports, financial reports or news articles. However, text-based methods for measuring innovation are still quite rare.

Bellstam et al. (2019) developed a text-based method for recognizing innovative firms. This method uses text analysis to cluster firms based on analyst reports and chooses the cluster with the most similar language to an innovation textbook. The method was found

to capture the innovativeness of companies that do not engage in R&D activities or patenting, but the results strongly correlated with patent data. What is more, this innovation measurement method was strongly correlated with valuable patents.

Mukherjee et al. (2017) also used a text-based method for measuring innovation. In this paper, the innovation measurement was based on new product announcements searched from a news database using specific keywords. The articles are compared with abnormal returns over a three-day period around the product announcement to filter out the articles that indicate major innovations. This method also takes into account innovations that are not patented or product launches made by firms without R&D activities. On the downside, it does not consider process innovations and minor or incremental innovations may be left unnoticed, if they do not induce a market reaction.

2.3 Hypothesis

In the study by Bellstam et al. (2019), innovative text in analyst reports was connected to firm innovativeness. Financial reports are different from analyst reports by content, but drawing from this evidence, relation to innovation obtained from other than analyst reports should be explored. The study itself is unprecedented, since no similar studies have been published, as far as is known.

Theoretical grounds of the hypothesis lie in the observation that the language and words used in financial reports correlate with firm characteristics, such as profitability and deceptive behaviour. This is supported by the studies of Patelli and Pedrini (2014) and Leung et al. (2015). This observation calls for studying other characteristics that can be inferred from financial reports' text data. The observation that innovation and language can be connected made by Bellstam et al. (2019), strengthens the assumption that innovativeness could be present and measured in the financial report text.

Holland (2009) writes that companies have market-based incentives to disclose value relevant information and it would only be logical for firms to describe their innovative activities in financial reports. As explained in previous chapters, innovations tend to increase firm value, and consequently, disclosing innovativeness should be profitable for the firm. Also, investors could be drawn to innovative firms due to higher expected returns, which is why firms should have an initiative to disclose their innovativeness. On the other hand, if firms that are not innovative, also have an incentive to seem innovative, distinguishing innovative firms from non-innovative based on their own disclosure becomes difficult.

H_1: *Innovativeness correlates to the language used in firms' financial reporting*

If a correlation between innovation and financial report language is found, a good basis for consequent research on the goodness of this measurement is formed. Further research on the text-based innovation indicator could be made to find out whether it measures innovation more comprehensively than traditional innovation indicators.

3 Text Analysis of Accounting Reports

Language and human-produced text can be used for quantitative analysis similarly to number data. Natural language processing (NLP) is a collective term for computational processing of human language with either the input or output of the algorithm being natural language such as human-produced text (Goldberg, 2017, p. xvii). Natural language processing is based on statistical machine learning, but unlike numbers, human language is changing and ambiguous, which makes it harder to analyse computationally. Natural language processing models can be supervised or unsupervised and linear or nonlinear, which is more closely examined in this chapter. In the first part, the process of an NLP-task is presented and the second part covers different methods of natural language processing.

When it comes to analysing corporate reports and more specifically, their narrative parts, there are various natural language processing methods that can be used to analyse the texts. Loughran and McDonald (2016) write that predicting firms' returns, bankruptcies or stock market fluctuations are all issues that could be answered by textual analysis. A text-analysis method could pick up patterns in, for example, Twitter posts or news articles that take humans long to apprehend. Or in this case, accounting reports. Reading, say, 1000 accounting reports takes a long time for the average person, but a computer can do this in seconds whilst conducting analysis and finding intricate patterns in the text.

Financial reports include qualitative content that does not represent numerical information. This sort of language information is less commensurable than pure numerical information, which is why text analysis methods can be helpful in finding valuable information from financial statements. According to Lewis and Young (2019), the numeric contents of financial statements do not contain nuances similar to verbal discourse and qualitative content in financial statements gives valuable information about the firm. Lewis and Young also report a significant increase in qualitative information in annual reports, the word count of firms listed in The London Stock Exchange had increased from 14,954 to 33,193 words over the period of 2003-2016.

If these narrative parts can be used to predict firm returns or share value, as mentioned above, maybe this information is useful for more distinct information too. This study aims to answer whether the narratives of financial reports can be used to determine if a firm is innovative.

3.1 Statistical Natural Language Processing

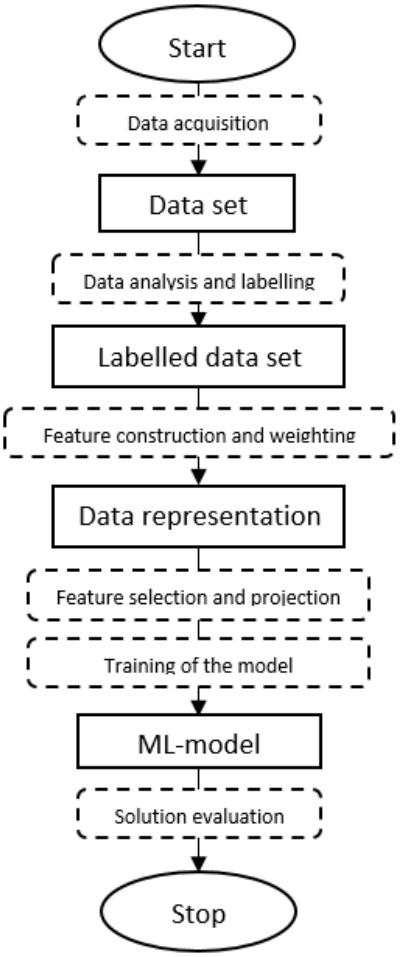

Figure 1 Machine learning process (Adapted from Mironczuk & Protasiewicz, 2018)

Figure 1 describes the machine learning process adapted from the text classification process defined by Mironczuk & Protasiewicz (2018). The process of text classification is similar to other NLP-methods and most of the phases are universal. The focus of this study is on text classification methods, since most text analysis problems are classification problems, but some other types of algorithms will be covered as well.

According to Mironczuk & Protasiewicz (2018), the process starts with data acquisition from the selected text source, which becomes the data set. To study the data, pre-processing is required to present the data correctly for the learning method to understand it. In text analysis, pre-processing can be e.g. tokenization or stemming. The feature construction and weighting phase comes after pre-processing, which continues to remodel the data into a form that the algorithm can use. Next, the features of the text need to be reduced and the dimensionality of the data needs to be lowered, so that only the necessary parts of the data for the analysis remain. Before model testing, the algorithm needs to be trained with a different data set from the one used for testing. Training is necessary so that the algorithm learns its target. If the training is successful, the algorithm should now be able to process incoming data similarly to the training data set. Finally, the evaluation of the model is required to assess its viability.

3.1.1 Text Pre-Processing

Because text in itself is highly dimensional and found in various mediums, pre-processing is needed before a machine learning algorithm can comprehend the text-data. Basically, text is qualitative data and to apply traditional- or text-analysis methods, it needs to be converted to a quantitative form (Loughran & McDonald, 2016). The pre-processing also includes removing words and the aspects of the texts that are not necessary for the analysis itself. The phases of text pre-processing include vector space model creation, feature selection and feature projection (Mironczuk & Protasiewicz, 2018).

Lemmatizing, stemming, lexical resources and distributions can be utilized in feature selection of single words without context (Goldberg, 2017, p. 67-69). In lemmatizing, the lemma of the word is used to combine similar words into their common lemma. In other words, the basic form of a word is used instead of the inflected form that appears in the original text. Stemming is another way of shortening words by their common letter sequences, in stemming, plurals, singulars and different tenses are shortened to one representation.

Lexical resources are dictionaries meant to be assessed by machines, and they include information about the meaning of a word and words which are similar to it (Goldberg, 2017, p. 67-69). Also, the distributions of different words can be used to find ones that behave similarly to extract their meanings. In stop-word removal, the words that hold no significance and are common to the documents, like *the*, *for* or *to,* are removed (Aggarwal & Zhai, 2012).

All the above mentioned feature selection models treat text to its linear order. Because language does not consist of a linear order of words and often contains difficult-to-observe features, complicated feature selection models also exist and can be used to infer linguistic properties, combination features, word sequences or distributional features (Goldberg, 2017, p. 70-76).

According to Enriquez et al. (2016), the bag of words –method (BOW) is the most frequently used method for text representation, more specifically, the BOW transforms text into sparse vectors. The bag of words generates a vector of the text, which can be a sentence, a paragraph or a document and the vector is based on a dictionary. Each word has an ID indicating its position in the vector. The weakness of the bag of words is that it does not account for word order and loses the semantics of the words (Le & Mikolov, 2014).

Word2Vec is a vector representation model, like the bag of words, created by Mikolov et al. (2013). It is a neural network -based skip-gram model that attempts to present the words in vectors useful for predicting the surrounding words. Each word is represented by a column in matrix *W* and the words can predict other words from calculations, such that “Berlin” – “Germany” + “France” should equal the word “Paris”. A shallow neural network is used to train the word vectors from a training dataset. Word2Vec can also capture the meanings of words as it maps words with similar meanings to similar vectors (Le & Mikolov, 2014).

Doc2vec is a vector representation model for representing entire text paragraphs or documents, whereas the models described previously only capture individual words or sentences (Le & Mikolov, 2014). Compared with the bag of words, the Doc2vec-model also attempts to capture semantics such that similar words would have more similar vectors. The paragraph vectors are unique vectors with common, fixed word vectors and an important feature to them is that they capture the word order. Doc2Vec uses a shallow neural network with one hidden layer. The difference of Doc2Vec compared with Word2Vec is that it adds a paragraph token to the output vector that represents the missing information regarding the context.

3.1.2 Model Training

A rough division of statistical learning problems is supervised and unsupervised learning (James et al., 2017, p. 26-28). According to Mironczuk & Protasiewicz (2018), in supervised learning, the data is pre-labelled for the algorithm with input and output values and basically the algorithm learns to make generalizations from the training dataset. According to Kirk (2017, p. 15-16), unsupervised learning is about the algorithm trying to understand the given data without feedback, for example clustering is an unsupervised learning method. The learning methods are discussed in detail in Chapter 3.2.

Before the actual dataset, the machine learning algorithm is presented with a training dataset, which it can use to learn the correct classification (Mironczuk & Protasiewicz, 2018). The test set data can never be used for model training and sometimes even three datasets could be used, one for training, one for validation and the final dataset for model testing and future error rate calculation (Witten et al., 2011, p. 149). Splitting the data for training, possible validation and testing is quite simple with a large dataset. The model needs enough data to make efficient generalizations, and therefore, when only a limited amount of data is available, splitting the dataset could become problematic. K-fold cross-validation, which is explained in the next chapter, is one of the possible solutions to too small a dataset.

3.1.3 Model Evaluation

To evaluate the classification algorithm performance, the examination of the training dataset classification outcome is not sufficient for model evaluation (Witten et al., 2011, p. 147-148). Model evaluation methods exist to evaluate the classification performance on the test set. The error rate of the training data is not a good indicator of performance on the test data because the performance estimation would be too optimistic. Information retrieval (IR) differs from classification or clustering because it has a lot of possible answers and IR models need different evaluation methods and indicators (Nakache et al., 2005). For example, document similarity measures fall into the information retrieval category.

According to Wong (2015), k-fold cross validation and leave-one-out cross-validation are common methods of evaluating a classification algorithm. K-fold cross-validation is suitable for a large dataset and leave-one-out cross validation for a situation with a limited amount of data. In k-fold cross-validation, the dataset is split into *k* random groups with similar class representations and the model is then trained on *k*-1 groups and tested on the hold-out group and this is repeated so that every group takes turns as the hold-out

group (Witten et al., 2011, p. 153). Lastly, the errors generated are averaged for an overall error estimate.

In leave-one-out cross-validation, the number of folds equals the number of instances and is thus suitable for a small dataset (Wong, 2015). Each instance is left out on its turn and the model is trained on the remaining instances (Witten et al., 2011, p. 154). The model is evaluated based on a successful classification of the hold-out estimate. Leave-one-out cross-validation does not involve random sampling and the process is not repeated at all, it is executed exactly *n* times. The error is formed from an average of the *n* judgments.

Other performance measures for evaluating the model after the classification are various statistical indicators such as precision, recall, F-score, error rate and area under the curve (Mironczuk & Protasiewicz, 2018). F-score is actually a combination of the precision and recall indicators, precision measures the proportion of positive identifications that were correct and recall measures the proportion of actual positives identified correctly (Nakache et al., 2005). According to Sokolova & Lapalme (2009), the classification success can be evaluated in four different ways; computing the number of correctly recognized class examples, correctly recognized examples that do not belong in the class, examples with an incorrect assignment to the class and examples belonging in the class but left unrecognized.

Thompson et al. (2015) measured the performance of textual similarity algorithms using recall as the performance measurement. First, the similarity algorithms were given the task of finding the most similar documents to the source text out of documents of different levels of plagiarism. Then recall was measured at different retrieval intervals compared with expected relevant documents. Cosine similarity had the highest recall for highly similar or heavily reviewed texts and the second highest for lightly reviewed and highly dissimilar texts.

3.2 Natural Language Processing Methods

This chapter presents some natural language processing methods used in analysing firm financial disclosure and reviews their applications in literature. The literature presented in this chapter studies data retrieved from either firm financial reports or analyst reports and analyses it with a certain text analysis method. Literature uses the term *narratives* when discussing the narrative sections of financial disclosures i.e. other than numeric disclosure.

According to Fisher et al. (2016), accounting and finance literature uses many different artificial intelligence (AI) and machine learning (ML) methods for NLP. For example different neural network methods, support vector machines and statistical classifiers are used. Some of these methods are presented in this chapter, including Latent Dirichlet allocation, which is the method used in this study.

3.2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is an unsupervised probabilistic classification model that was first introduced by Blei et al. (2003) to identify topics in a large text corpus, unlike most of the other statistical models used for NLP, it was developed specifically for text processing. According to Dyer et al. (2017), the LDA compares the probabilities of different words occurring in documents to assign the documents to latent topics. After the LDA has identified different topics, the researcher assigns labels to the topics. Due to being unsupervised, researcher bias does not affect the LDA results, although the LDA does need researcher help in narrowing down the number of topics for the sake of interpretability.

The LDA allows for multiple topics and can distinguish different topics in the same corpus, which is why it is well suited for 10-K documents (Dyer et al., 2017). 10-K documents

contain different topics depending on the narrative section and one firm can fall into multiple classes.

Bellstam et al. (2019) used the LDA in their research of corporate innovation to classify corporations to different topics based on analyst reports and finding the topic where innovative corporations were classified (see Chapter 2.2.4). Dyer et al. (2017) also used the LDA in their study to identify topics in 10-K reports. LDA is also used in this study, similarly to Bellstam et al. (2019).

3.2.2 Support Vector Machine

The support vector machine was first introduced by Cortes and Vapnik (1995) and has since become a popular text classification algorithm. SVM is a linear classifier that can be used for the classification tasks of both linear and nonlinear data (Onan et al., 2016). Support vector machine is a learning algorithm that projects the data into a multi-dimensional hyperplane and draws the partition boundaries. It tries to draw an optimal partition line to classify the data into two classes.

Because text data is usually high-dimensional, the SVM can simplify the classification with its ability to draw a decision boundary between the classes (Kirk, 2017, p. 110). The decision boundary (hyperplane) is drawn so that the margin ξ between the classes is maximized (Allahyari et al., 2017). Those text vectors that lie at a ξ distance from the hyperplane are called support vectors. In the case of data that is not linearly separable, the SVM classifies the data while trying to minimize the number of vectors on the wrong side.

Chen et al. (2017) used a support vector machine method to detect fraud in narrative reports. Humphreys et al. (2011) also tried to identify fraudulent statements using textual analysis methods, they compared the results from multiple methods, including the

SVM, Naïve Bayes and logistic regression. Naïve Bayes performed with the highest overall accuracy out of these three but with small differences. Purda and Skillicorn (2014) classified financial reports as fraudulent and non-fraudulent based on predictive words with the help of a SVM.

3.2.3 Neural Networks

A neural network is a kind of mathematical representation of the brain; in a neural network, a neuron is one computational unit (Goldberd, 2017, p. 41). A neuron receives a vector of inputs and each neuron has a certain set of weights that it uses to compute a function with its inputs (Aggarwal & Zhai, 2012). A feed-forward neural network typically consists of layers, the bottom layer is the input layer and the top layer is the output layer, between them lie middle layers, which represent a nonlinear function (Goldberg, 2017, p. 41-42). If all the neurons in one layer are connected to all the neurons in the next layer, it is a fully connected layer. The layers of a neural network are actually vectors and between the layers, linear transformations are performed on the vectors. The neural network can be utilized to, for example, regression, binary classification and k-class classification.

The main types of neural networks are feed-forward networks and recurrent networks (Goldberg, 2017, p. 3). Feed-forward neural networks are good at extracting patterns in the text and identifying indicative phrases. Convolutional networks are a special type of feed-forward networks, where there are multiple deep layers, one of which is a convolutional layer. The convolutional layer finds local connections and relationships from the previous layer. Recurrent neural networks on the other hand, are specialized in sequential data, and they take a sequence of items as an input, of which they summarize a sequence. Recurrent neural network outputs can be used as inputs to a feed-forward network.

Matin et al. (2019) used a neural network model to predict a corporate distress probability using annual report text segments. The study was conducted by extracting patterns from the text with a convolutional (feed-forward) neural network and feeding its output to a recurrent neural network for pattern understanding. Finally, with the help of numerical financial variables, a probability of distress is predicted. Rönnqvist and Sarlin (2017) studied bank distress from bank distress events and the language of news data with the help of a neural network.

3.2.4 Statistical Classifiers

I. Regression Classification

Regression methods commonly applied to numerical data can also be used for text classification. The Linear Least Squares Fit (LLSF) is a regression method for text classification (Aggarwal & Chai, 2012, p. 196). LLSF categorization was first introduced by Yang and Chute (1994). The LLSF makes the categorization based on a human-categorized training sample, which is called “example-based relevance judgments”. The goal of the LLSF is to minimize

$$\sum_{i=0}^{n}(p_i - y_i)^2, \quad (1)$$

where p_i is the predicted class label and y_i is the real class label.

Logistic regression is a classification algorithm that is most typically used for binary classification, where the target value is between 0 and 1 (Onan et al. 2016). Logistic regression models the probability of an event as a linear function of the predictor variables. It is similar to linear regression, but linear regression is unable to capture probabilities and hence does not produce values that are usable for estimating probabilities. According to

Witten et al. (2011, p. 125-126), the least-squares assumes that the errors are statistically independent and normally distributed with the same standard deviation, which is not possible when the observations take the values of 1 or 0.

Kim and Kim (2014) regressed the investor sentiment index with stock returns in their study, however, they used Naïve Bayes for labelling the messages studied as “buy” or “sell” to define an index for sentiment. Tsai and Wang (2017) used regression methods to study the ability of soft information in financial reports to predict firm risk.

II. K-Nearest Neighbours

K-nearest neighbour (KNN) classifier makes an estimate for the conditional distribution of Y given X and classifies the observation to the class with the highest probability (James et al., 2017, p. 39). When we have a test observation x_0, the KNN classifier will identify the K closest points to the test observation and assign it to the class with the highest probability amongst the K points.

The similarity measure used in the KNN could be the number of common words in the documents with normalized document lengths (Hotho et al., 2005). Words have varying information content and there are other methods that also account for this, such as cosine similarity. In the vector space model, the documents are represented by a numerical feature vector (Groth & Muntermann, 2011). The similarity of the vectors can be compared, e.g. by Euclidean distance. The high dimensionality of textual data can be a complication in using KNN for text classification (Kirk, 2017, p. 110).

The KNN among other machine learning methods was used by Groth & Muntermann (2011) to study the effects of corporate disclosures on risk. Huang and Li (2011) used a multilabel categorical KNN to extract risk factors from 10-K filings.

III. Decision Trees

A decision tree is built with recursive binary splitting until a sufficient tree is formed (James et al., 2017, p. 311). In a classification tree, we assume that each observation belongs to the class that is the most commonly occurring in the training sample. Gini impurity, information gain and variance reduction are common methods for splitting data into subcategories (Kirk, 2017, p. 71-73). How information gain works, is that it finds the attributes that improve the model and makes a split at those points. Gini impurity is calculated as a probability of a factor appearing in a given class and the first split point is chosen by the least impurity and thus the highest probability of a correct classification. Variance reduction can be used for continuous trees, and it aims to reduce the scattering of the classification.

The Random forest is an application of the decision tree, which is constructed of multiple decision trees and the output is the statistical average of the decision trees (Heller, 2019). The Randomness comes from the forest being constructed by using bagging for taking a random subset of features for each decision tree to eliminate the effect of a single very strong decision point.

Decision trees are not very commonly used for text analysis in accounting and finance literature. Wang et al. (2013) used a decision tree to investigate the effect of the contents of an information breach announcement on stock prices. Henry (2006) studied the effect of verbal predictor variables in earnings press releases on market performance using a tree-based algorithm.

IV. Naïve Bayes Classifier

Naïve Bayes (NB) classifiers are a family of probabilistic classifiers and they are likely the simplest text classification models (Xu, 2018). What makes the classifier naïve is that it assumes that all features are independent of each other. According to Xu, NB is quick and easy to implement and works well with text classification, which is why it can be

used as a baseline in text classification. According to Dib & El Hindi (2017), NB is simple and practical, which is why it is one of the best performing algorithms.

In text classification, Bernoulli Naïve Bayes and Multinomial Naïve Bayes are typical NB methods (Diab & El Hindi, 2017). In Bernoulli Naïve Bayes each document is a vector of binary numbers where the presence of a word is indicated as 1 and absence as 0. Multinomial Naïve Bayesian represents a document as a vector of words and labels it based on the count of these words in the document.

Naïve Bayes classification was used by Besimi et al. (2019) to predict stock price fluctuations caused by financial news. The classifier was given articles with negative or positive sentiment and tasked with classifying the market reaction as either up or down. Huang et al. (2014) used Naïve Bayes classification to analyse the sentiment of analyst reports. They assign sentences to classes by their sentiment with a Naïve Bayesian and compare these results with abnormal market returns. Buehlmaier and Whited (2018) used the NB to analyse firms' financial constraints, but on the contrary to the former studies, they used the NB to produce a probability of financial constraint instead of pure classification.

3.2.5 Textual Similarity

Semantic textual similarity (STS) is a natural language processing tool that measures and scores sentences based on their similarity (Lopez-Gazpio et al., 2017). STS is a measure of semantic similarity between documents and it consists of direct and indirect relationships measured through their semantic similarities (Majumder et al., 2016). The similarity is then graded at a scale of 0 to 5, 0 being not at all similar and 5 being completely similar.

STS is a general concept of measuring similarities between texts and it includes different methodologies, such as topological, statistical and string-based methods (Majumder et al., 2016). Topological methods include node-based, edge-based and hybrid models, all

of these methods consider semantic relationships between the words. In statistical similarity, a statistical model is built before the similarity is estimated, Latent Semantic Analysis is an example of statistical similarity measures. Cosine similarity, which is introduced in the next chapter, is a string-based textual similarity measure.

Kamaruddin et al. (2015) developed a text mining system for detecting deviations in financial documents, because classification was insufficient for this task and did not provide the tools for textual comparisons and semantic analysis. Their method gives a similarity or dissimilarity score to the studied text and was deemed efficient at this task.

Similarity measures differ from other methods presented in this chapter so far, because they are not classification models but measure the similarity between sentences or documents. Cosine similarity is also a measure of textual similarity, but it is measured for numeric vectors from the text (Goldberg, 2017, p. 119). Cosine similarity is computed by measuring the cosine of the angle between the vectors:

$$sim_{\cos}(\boldsymbol{u}, \boldsymbol{v}) = \frac{\boldsymbol{u} * \boldsymbol{v}}{\|\boldsymbol{u}\|_2 \|\boldsymbol{v}\|_2} \quad (2)$$

Cosine similarity can be measured from word or document vectors generated, e.g. with word2vec or doc2vec. The similarity measure it returns ranges between 1 and -1, 1 being exactly the same and -1 exactly opposite. The 0 value indicates decorrelation (Park et al. 2020). Cosine similarity also accounts for document length by normalizing the text vectors (Hoberg & Phillips, 2016). Hoad and Zobel (2003) found that cosine similarity measure performs the best at retrieving the most similar documents when they are of varying length or distinctly different from the rest of the corpus.

The cosine similarity measure was used by Hoberg and Phillips (2016) to evaluate similarities in 10-K product descriptions. They used the words that firms used in their 10-K

product descriptions and mapped them into industries based on a pairwise cosine similarity of the words. Peterson et al. (2015) studied firms' accounting consistency by measuring the cosine similarity of the accounting policy disclosures.

4 Data

The data used in this study are 10-K reports retrieved from the U.S. Securities and Exchange Commission (SEC) EDGAR (2020) database. The EDGAR database lists U.S. public companies' annual, quarterly and current reports, among other reports (U.S. Securities and Exchange Commission, 2018). This study uses annual reports, which are found under the name 10-K in the database and correspond to annual reports. The sample used for this study includes firms from various industries across the years 2008-2018. The sample size for training the model is 45971 individual 10-K documents and after matching the firms in the sample with available financial and patent data, 8364 firm-year observations are used for validating the method. The methodology used in this study is described in detail in chapter 5.

In addition to 10-k filings, innovation text samples were used to construct the innovation measure. Eight different text samples were chosen to test, which of them correlates the most with known innovation measurements (patents and R&D expenditure). All of the texts are chapters of innovation textbooks following Bellstam et al. (2019). Also, one text sample of a corporate finance book was taken as a control for the innovation texts. A list of the texts can be found in the appendix.

For validating the innovation measure, corporate financial data was used. The financial variables used were the balance sheet value of firms' patents and brands, and research and development expenditure as response variables and total assets, net sales or revenues, total liabilities and return on assets as control variables.

4.1 Form 10-K

The form 10-K offers detailed information about the company's business, risks, financial result and the fiscal year (U.S. Securities and Exchange Commission, 2011). U.S. public

companies are required to file a 10-K form to the U.S. Securities and Exchange Commission yearly. Financial reports are an important source of information for investors due to the broad range of information they offer. The form 10-K annual report includes 15 items in four parts but one company might not need to disclose all items if they do not concern said company.

Part I includes items 1 “Business”, 1A “Risk Factors”, 1B “Unresolved Staff Comments”, 2 “Properties”, 3 “Legal Proceedings” and 4, which is reserved for future rulemaking but does not have required information as of the moment (U.S. Securities and Exchange Commission, 2011).

Part II includes items 5 “Market for Registrant’s Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities”, 6 “Selected Financial Data”, 7 “Management’s Discussion and Analysis of Financial Condition and Results of Operations”, 7A “Quantitative and Qualitative Disclosures about Market Risk”, 8 “Financial Statements and Supplementary Data”, 9 “Changes in and Disagreements with Accountants on Accounting and Financial Disclosure”, 9A “Controls and Procedures” and 9B “Other Information” (U.S. Securities and Exchange Commission, 2011).

Part III includes items 10 “Directors, Executive Officers and Corporate Governance”, 11 “Executive Compensation”, 12 “Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters”, Item 13 “Certain Relationships and Related Transactions, and Director Independence” and 14 “Principal Accountant Fees and Services” (U.S. Securities and Exchange Commission, 2011).

Finally, Part IV includes item 15 “Exhibits, Financial Statement Schedules” (U.S. Securities and Exchange Commission, 2011).

4.2 Other Data

The data used in this study is text data from 10-K filings, as mentioned in chapter 4.1, but also text data from innovation textbooks and numerical data in the form of the firms' financial data. The other text data source was innovation text samples, used to define innovative topics and measure the innovativeness present in a certain 10-K report. The innovation text samples are samples extracted from different innovation textbooks but some of the samples are from the different parts of the same book, a full list of the texts can be found in the appendix. Some of the text samples are from the introduction chapter, because it is expected that this chapter would have the most general innovation language, but samples from other parts of the books are included too.

To prepare the text data for analysis, numerical information, special characters, email addresses, websites and words of only one character were removed. The text was also tokenized and lemmatized, stop-words were removed and all text was lowercased.

The form 10-K includes detailed information about the company's key business and main products. Also, research and development activities are usually disclosed in these forms. It is expected that the form 10-K includes innovation-related disclosure in the form of business activities, product information and R&D activity.

Other data used to validate research results are balance sheet values of patents and brands, R&D expenditure and other firm-level financial data. All these key figures are from the Refinitiv Eikon (2020) database. R&D and patents and brands are presented as a percentage of the firms' net sales or revenues in this study.

4.3 Descriptive Statistics

Table 1 shows descriptive statistics on the lengths of the text documents. In the first column are the statistics for the full sample of 10-K documents and in the second column

innovation texts. The 10-K filings have a varying length from 195 words to 1 190 370 words, whereas the innovation texts' word length varies less.

Table 1 Descriptive statistics on document length

	10-K	**Innovation texts**
Mean	60 752.64	10 390.88
Median	51 148.00	11 724.00
Min.	195.00	1 910.00
Max.	1 190 370.00	18 094.00

Figure 2 shows the most common words in the 10-K reports of the year 2016 in the sample when stop words are removed. The most common words are similar for the rest of the years as well, but due to slow computation, only one year could be taken into inspection at once. All of the words are quite expected for financial reports. The word "company" is by far the most used word in the documents, which is not very surprising.

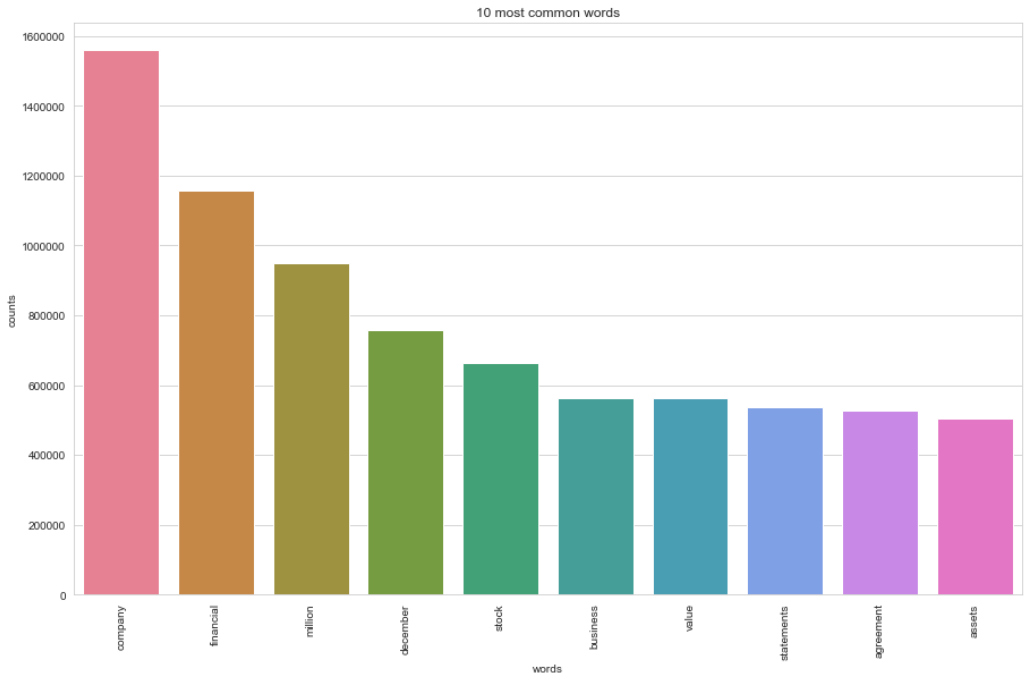

Figure 2 Most common words in 10-K filings in 2016

The financial data was winsorized before conducting the regression analysis, in this case the bottom 1% and top 5% of the observations were removed due to the skew in the distributions of research & development and patent variables. Research & development and patent values are both presented as a percentage of sales to control firm size. In table 2 are descriptive statistics on all of the financial variables used in this study.

The firm-year financial variables are explained in detail in chapter 5.3, where the regression models are presented. In short, the variables in table 2 are the following:

$\frac{P_{it}}{sales_{it}}$ represents the balance sheet value of patents and brands as a percentage of sales. $\frac{RD_{it}}{sales_{it}}$ represents research and development spending as a percentage of sales. $\log(TA_{it})$, $\log(sales_{it})$ and $\log(TL_{it})$ are the logarithms of total assets, net sales or revenues and total liabilities. ROA_{it} represents return on assets and $\log(sales_{it+1})$ represents the logarithm of the growth or decrease in sales from t to t+1.

Table 2 Descriptive statistics on financial data

	$\frac{P_{it}}{sales_{it}}$	$\frac{RD_{it}}{sales_{it}}$	$\log(TA_{it})$	$\log(sales_{it})$	$\log(TL_{it})$	ROA_{it}	$\log(sales_{it+1})$
Mean	7.81	10.39	13.35	13.06	12.48	-3.52	0.054
Std.	12.03	14.19	2.13	2.31	2.40	27.70	0.36
Min.	0.02	0	7.84	5.85	6.96	-163.32	-5.57
25%	0.87	1.3	11.93	11.72	10.68	-3.30	-0.04
50%	2.82	4.66	13.46	13.33	12.63	4.58	0.05
75%	8.48	14.07	14.91	14.70	14.33	9.05	0.15
Max.	47.60	57.65	16.99	16.63	16.45	18.04	3.70

4.4 Data Issues

The data used in this study are public companies' annual reports. These reports are made by the company itself and thus are not objective reports. The companies' objective is to

attract investors and increase market value, which is why there exists an incentive to make the company seem as good as possible in the eyes of an investor. The main issue with the data is that companies trying to seem innovative might be practising impression management rather than being actually innovative. It needs to be taken into account in the study that the reports might be biased and the research results must be validated in a way that is independent of the companies' own disclosure. This chapter presents the main research regarding and ways of impression management and bias in financial reports.

4.4.1 Impression Management

Firms publish their financial statements in an annual report and the annual report contains financial information and accounting narratives, such as the President's Letter and Management Discussion and Analysis (Jones, 2010, p. 97-98). Firms trying to manage the presentation of the annual report is called impression management. Impression management can also be practised via graphs or photographs. Accounting narratives are not audited, which is why they are especially suitable for impression management.

According to Jones (2010, p. 99), there are four main impression management methods, which are stressing the positive and downplaying the negative, baffling the readers, differential reporting and attribution. Firms may more eagerly report positive news but leave bad news without mention. Also, bad news could be reported with a complicated language to baffle the readers, but with good news, a simpler language is used. Profitable and unprofitable firms might have different reporting strategies; for example, profitable firms might disclose more concrete accounting information to prove their superiority. Lastly, firms may take credit from good news but blame bad news on the environment. Merkl-Davies and Brennan (2007) suggested that impression management in financial reports questions the quality and usability of these reports to investor decision-making.

Cho et al. (2010) found that corporations with worse environmental performance used a biasing language in their environmental reports. The firms with worse environmental performance emphasized the good news while trying to cover the issues to give an impression of a better environmental performance. This raises the question whether this kind of impression management could be possible in terms of reporting innovativeness and are research results based on qualitative information in financial reports biased.

On the other hand, Patelli and Pedrini (2014) discovered that optimistic tone in the CEO letters of Fortune 500 firms was significantly associated with better financial performance. Firms with optimistic letters predicted better future returns and the CEO letters of the highest-earning firms were the most optimistic. This finding indicates that organizational communication is legitimate and shows no manipulation.

Leung et al. (2015) concluded that firms which practised minimal narrative disclosure also performed poorly and had a higher risk of financial distress. According to the paper, these firms were trying to withhold information about the firm and practised impression management.

According to the findings of Lobo et al. (2018), firms with more innovations have a lower financial reporting quality. This is assumed to be due to more innovative firms having more agency problems and a higher incentive for earnings management. The findings are also consistent with the hypothesis that higher audit quality mitigates the lowering effect of innovation on financial reporting quality. This behaviour might not be impression management strictly speaking, but does imply variability in disclosure, which in turn could have an effect on research results.

4.4.2 Signalling Theory

Signalling theory is based on the hypothesis that if a market has “good” and “bad” products and the value of the good products is higher from the bad products, but it is difficult

to distinguish the two products from each other, the market price will set at the price of the bad product (Dixit et al., 2015, p. 294-298). The sellers of the good and valuable products do not want to sell their product below its value, so they need to have a strong enough signal that the sellers of the bad products cannot replicate the signal, to show that their product is the good one.

According to Moratis (2018), signalling theory includes four key theoretical concepts; signals of quality and intent, the efficacy of signalling, signal honesty and fit and signal frequency and consistency. Signals of quality indicate a certain organizational characteristic, whereas the signals of intent indicate a future action. Observability alone does not fill the condition of signal efficacy, the signal also needs to be costly. The signal needs to correlate with the unobservable quality of the signaller to be deemed fit and honest. To increase the effectiveness of a signal, firms can increase the signal frequency or use multiple signals for the same message.

Janssen and Roy (2013) argue that firms have information that is not publicly available about the quality of their products, and they have the opportunity to voluntarily disclose the private information verifiably. However, Jansen and Roy argue that firms do not engage in this voluntary disclosure but signal their quality through market activities, like pricing. Innovativeness as a whole is more complicated than launching a new product, but new products are a dimension of innovation and one way of signalling innovativeness could be launching innovative products.

Drawing from the logic of signalling, firms would not have incentives to pretend to be innovative without verifiable signals to prove it. Innovation outputs are a way of proving the firm innovative, but an interesting question is whether these innovative firms also use language and words in their reporting that separates them from firms that are less innovative.

5 Methodology and Research Design

The research methodology roughly follows the study by Bellstam et al. (2019). The procedure goes as follows: first, a LDA-model is trained with the 10-K filings and then used to extract topic distributions from innovation texts. Secondly, Kullback-Leibler divergence is calculated between each firm-year-filing and innovation text and used as the innovation measurement. Lastly, the measurement is validated by regressing it with patent- and R&D-values and the effectiveness of different innovation texts is evaluated based on these results.

5.1 LDA

Latent Dirichlet allocation, as described in chapter 3.2.3, was used in this study to define the topics that innovative firms use in their 10-K filings. The model was constructed to allocate k=15 different topics. Due to memory restrictions, the model was trained on a sample consisting of 10-K filings from the same year and then the model was updated one year at a time on the rest of the samples, approximately 4000 documents at a time.

Table 3 shows the 15 topics and the 10 most common words in each topic. Most of the words are very general to financial reports, for example, *million, financial, company, tax, asset* etc. But, the topics also include almost company-specific words, like topic 6, where *entergy* and *louisiana* are among the most common. Other topics clearly differing from others are topic 5, which includes many oil-industry related words, topic 8, which has many loan-related words, topic 10, which includes real-estate related words and topic 12, which has medical industry -related words. In most of the topics clear company-, industry- or subject-related words are not really distinguishable in the 10 most common words.

Table 3 The most common words in LDA topics

Topic	10 most common words
0	partner, unit, partnership, agreement, general, million, financial, cash, service, december
1	million, company, financial, asset, tax, cost, sale, product, value, fiscal
2	company, share, stock, common, business, note, director, financial, security, interest
3	service, financial, health, year, state, program, million, december, revenue, result
4	company, financial, year, statement, product, fiscal, control, stock, report, ha
5	gas, oil, natural, price, reserve, production, cost, property, well, financial
6	entergy, cost, louisiana, corporation, financial, million, system, texas, new, nuclear
7	energy, company, cost, power, million, financial, rate, gas, statement, asset
8	agreement, party, section, agent, lender, date, loan, respect, term, borrower
9	loan, financial, company, bank, interest, december, million, value, asset, rate
10	property, lease, million, llc, december, financial, tenant, company, interest, year
11	revenue, service, financial, customer, million, business, result, product, tax, could
12	product, clinical, development, patent, company, trial, candidate, agreement, u, drug
13	plan, company, participant, executive, agreement, section, date, employee, award, benefit
14	investment, financial, company, million, loss, value, income, december, insurance, risk

After training the model with all the 10-K filings, the model was used to extract topic distributions from innovation texts, which were previously unseen by the model. The topic distributions found in the innovation texts are seen in the results chapter 6.1.

5.2 Kullback-Leibler Divergence

Kullback-Leibler divergence (KL-divergence) was first introduced by Kullback and Leibler (1951) and it is a distance measure for probability distributions. Following Bellstam et al. (2019) and Lowry et al. (2019), KL-divergence was chosen as the distance measure for LDA-topics in this study to construct the innovation measure.

The innovation measure was constructed by computing the KL-divergence of the topic distributions of the 10-K filings and the innovation texts. A lower KL-divergence indicates higher similarity and thus the 10-K's with the lowest KL-divergence with the innovation texts are the documents with the most similar topic distributions to the innovation text. To be able to observe the relationship between the increase in the similarity between the texts and an increase in the other variables, the KL-divergence was multiplied by -1 for the regressions. Differing from the studies by Bellstam et al. (2019) and Lowry et al. (2019), in which a single innovation topic was chosen and its word distribution was then compared to other topics, this study uses the topic distributions of individual documents to compute the innovation metric. Multiple topics were perceived prevalent in the innovation texts and thus it seemed more natural to use the topic distributions to find out which 10-K filings had the highest prevalence of these topics. The topic distributions are more closely inspected in chapter 6.

5.3 Regression Models

Following Bellstam et al. (2019), regression models are formed to validate the results and study the effectiveness of the generated innovation measure. It is important that the new measure is correlated with these common innovation measures, to ensure that it is, in fact, effective in measuring innovation. As stated in the previous chapters, there are innovations that are not shown by the traditional innovation indicators and it is hoped that the new measure would also capture this kind of innovation. Nevertheless, there is a requirement that the measure captures "obvious" innovativeness, such as high patent value and R&D-expenditure.

Ordinary least squares (OLS) regressions were used for validating the innovation measurement. Linear regression models, where y is presented as a function of $(x_1, x_2, \dots, x_n)$ are formed for different combinations of variables to observe the effect of a change in an x-variable on y.

The following regression functions were used:

$$\frac{RD_{it}}{sales_{it}} = \beta_0 + \beta_1 innscore_{it} + \beta_2 \frac{P_{it}}{sales_{it}} + \beta_3 \log(TA_{it}) + \beta_4 \log(sales_{it}) + \beta_5 \log(TL_{it}) + \beta_6 ROA_{it} + X_{it} + \varepsilon_{it} \quad (3)$$

Regression function 3 observes the relationship between the innovation measurement and research and development. Out of the regression variables, $\frac{RD_{it}}{sales_{it}}$ acts as the dependent variable and represents research and development costs for firm i at year t as a percentage of sales for firm i at year t. $innscore_{it}$ is the variable representing the text-based innovation measure. The regression is also repeated eight times for each innovation measurement from different text samples. $\frac{P_{it}}{sales_{it}}$ represents the gross value of patents and brands for firm i at year t as a percentage of sales for firm i at year t, used to control for the relationship between R&D and patents. Control variables $\log(TA_{it})$, $\log(sales_{it})$, $\log(TL_{it})$ and ROA_{it} used to control firm size and characteristics following Bellstam et al. (2019), represent total assets, net sales or revenues, total liabilities and return on assets for firm i in year t, respectively. Logarithmic transformation has been made for the variables representing total assets, net sales or revenues and total liabilities. X_{it} represents the 10 industry dummies (1-9 and "none"), which are based on one-digit SIC-codes, also used as control variables. ε_{it} represents the error term.

$$\frac{P_{it}}{sales_{it}} = \beta_0 + \beta_1 innscore_{it} + \beta_2 \frac{RD_{it}}{sales_{it}} + \beta_3 \log(TA_{it}) + \beta_4 \log(sales_{it}) + \beta_5 \log(TL_{it}) + \beta_6 ROA_{it} + X_{it} + \varepsilon_{it} \quad (4)$$

Regression function number 4 studies the relationship between the innovation measurement and patents. The function has the same variables as function 3, but $\frac{P_{it}}{sales_{it}}$ now acts as the dependent variable and $\frac{RD_{it}}{sales_{it}}$ is one of the control variables, used to control the relationship between R&D and patents. This regression is also repeated eight times for each $innscore_{it}$-variable.

$$\frac{RD_{it}}{sales_{it}} = \beta_0 + \beta_1 CFscore_{it} + \beta_2 \frac{P_{it}}{sales_{it}} + \beta_3 \log(TA_{it}) + \beta_4 \log(sales_{it}) + \beta_5 \log(TL_{it}) + \beta_6 ROA_{it} + \varepsilon_{it} \quad (5)$$

Regression function number 5 acts as a control regression, it observes the relationship between the control innovation measurement and research and development. The function has the same dependent variable as function number 3, but with $innscore_{it}$ replaced with $CFscore_{it}$, which is the control firm-year innovation score derived from a corporate finance text sample to confirm that the effect of $innscore_{it}$ is specific to innovation text. Control variables $\frac{P_{it}}{sales_{it}}$, $\log(TA_{it})$, $\log(sales_{it})$, $\log(TL_{it})$ and ROA_{it} have the same definitions as in function number 3.

$$\frac{P_{it}}{sales_{it}} = \beta_0 + \beta_1 CFscore_{it} + \beta_2 \frac{RD_{it}}{sales_{it}} + \beta_3 \log(TA_{it}) + \beta_4 \log(sales_{it}) + \beta_5 \log(TL_{it}) + \beta_6 ROA_{it} + \varepsilon_{it} \quad (6)$$

Regression function number 6 has the same variables as function number 5, but $\frac{P_{it}}{sales_{it}}$ as the dependent variable and $\frac{RD_{it}}{sales_{it}}$ as a control variable instead.

$$salesgrowth_{it+1} = \beta_0 + \beta_1 innscore_{it} + \beta_2 \frac{RD_{it}}{sales_{it}} + \beta_3 \frac{P_{it}}{sales_{it}} + log(\text{sales}_{\text{it}}) + log(\text{growth}_{\text{it}-1}) + \text{X}_{\text{it}} + \varepsilon_{it} \quad (7)$$

Regression function number 7 observes the effect of the text-based innovation measurement on the growth of sales. In function 7, the dependent variable $salesgrowth_{it+1}$ represents the percentage change of sales for firm i from year t to t+1. The control variables $\frac{RD_{it}}{sales_{it}}$ and $\frac{P_{it}}{sales_{it}}$ are as explained in functions 3 and 4 and this regression is also repeated eight times for each $innscore_{it}$ variable. Other control variables include $log(\text{sales}_{\text{it}})$ and X_{it}, which are the same as in function 3, and, $log(\text{growth}_{\text{it}-1})$, which represents the logarithmic growth of sales for firm i from year t-1 to year t.

6 Results

This chapter presents the research results. In subchapter 6.1, the LDA topic distributions are inspected more closely and the firms with the top innovation scores are presented. In subchapters 6.2-6.5, regression results for functions 3-7 are presented and analysed.

6.1 Topic Distributions

Figure 3 shows the topic distributions for the different innovation texts. Topics 11 and 12 are clearly the most commonly occurring in almost all of the text samples. Topics 1, 3 and 4 are also quite common in all texts, whereas topics 0, 2, 6, 7, 8 and 9 occur very little in the innovation texts. Topic 12 included medical-related words (see ch. 5.1), but also the words “product”, “development” and “patent”, which should be common words for the innovation texts. According to Antonipillai et al. (2016), the medical industry is patent-intensive, and thus, if measured by patents, medical industry corporations should be generally more innovative than average. On the other hand, this could affect the innovation metric for other than medical companies, since, if they do not use the medical words in topic 12, their KL-divergence could be larger.

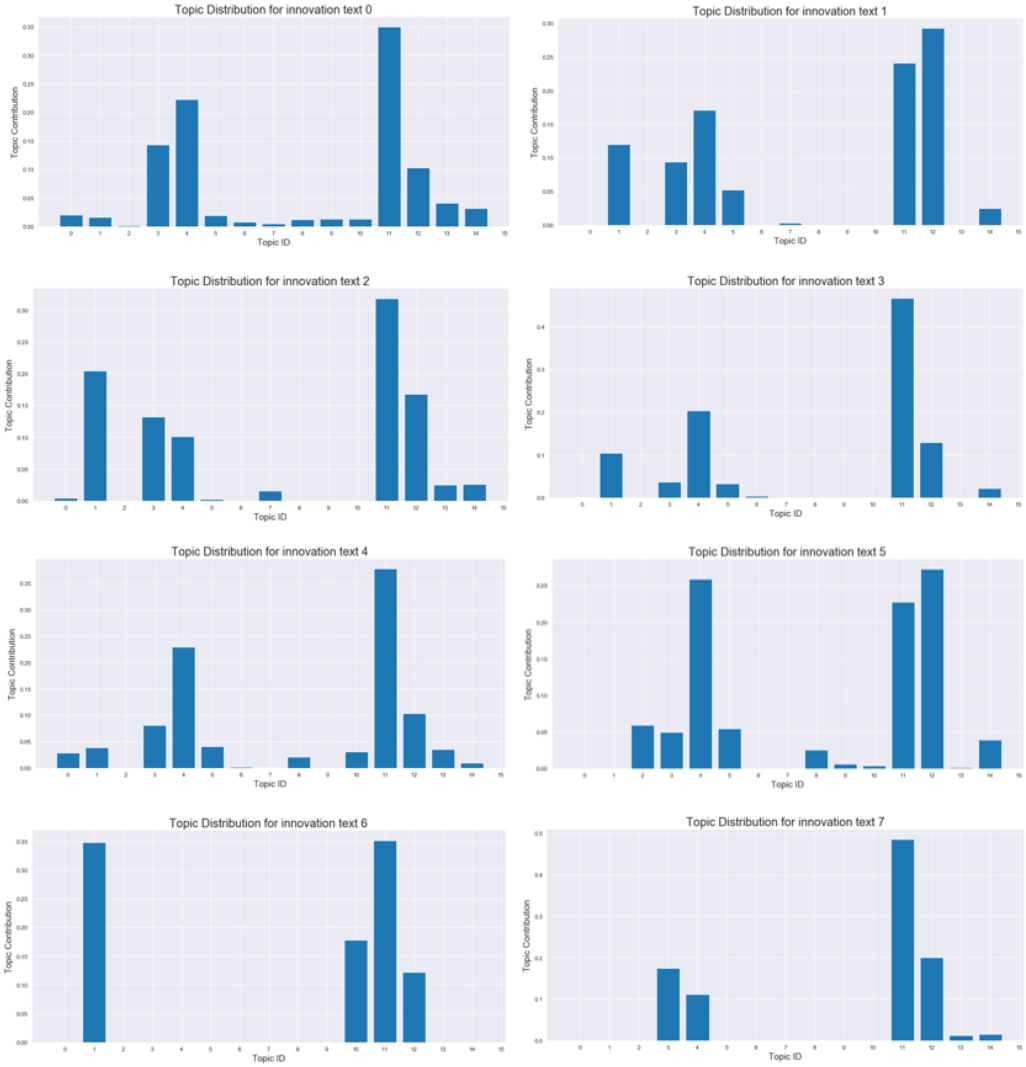

Figure 3 Innovation text topic distributions

The 10 firms from the sample with the best innovation score on average and their industry are listed in table 4. Most of the firms in the top 10 are from the fields of life sciences and manufacturing. According to Antonipillai et al. (2016), Medical equipment and supplies, Pharmaceutical and medicines, and technological industries are patent intensive. Also in the top 50 trademark registering companies, 28 represented other miscellaneous manufacturing and 28 pharmaceutical and medicine manufacturing and can thus be interpreted also as trademark intensive. A worry was expressed before that the method would measure innovative firms in the medical industry as more innovative than others. 5 out of 10 of the firms have an industry classification of “Office of Life Sciences”, but

there are many other industries in the top rating list. This result does not prove that the medical industry would be overrepresented compared to its innovativeness, but to control industry-specific effects, the industry dummies are added to the models.

Table 4 The firms with the highest innovation scores

Firm name	**Industry**
Axon Enterprise Inc.	Office of Manufacturing
Luna Innovations Inc.	Office of Trade & Services
Masimo Corp	Office of Life Sciences
Biolase Inc.	Office of Life Sciences
Tessera Technologies Inc.	Office of Manufacturing
Irobot Corp	Office of Manufacturing
Orchid Cellmark Inc.	Office of Trade & Services
Inogen Inc.	Office of Life Sciences
Quidel Corp	Office of Life Sciences
Cutera Inc.	Office of Life Sciences

Statistics on the KL-divergences of those 8364 firm-year observations included in the regressions are presented in table 5. The statistics in table 5 are for the original KL-divergences, but as mentioned previously, for clarity, the divergence was multiplied by -1 to form the regression variables and all of the values are negative in the regressions. Each "KL [No.]" in the columns of table 5 represents the divergence between the topic distributions of a certain innovation text sample and the 10-K filings.

Table 5 KL-divergences between each innovation text sample and the 10-K filings

	KL 0	KL 1	KL 2	KL 3	KL 4	KL 5	KL 6	KL7
Mean	2.072	2.522	2.017	1.890	2.037	2.513	2.534	2.569
Std.	1.041	1.464	1.225	1.164	0.988	1.322	1.327	1.525
Min.	0.123	0.150	0.125	0.160	0.195	0.165	0.294	0.219
25%	1.382	1.485	1.218	1.115	1.396	1.556	1.69	1.524
50%	1.863	2.197	1.736	1.639	1.847	2.206	2.247	2.235
75%	2.538	3.179	2.560	2.342	2.441	3.091	3.012	3.252
Max.	11.756	13.080	12.777	13.33	11.821	11.858	2.573	13.475

In table 6, there is a correlation matrix of the topic distribution KL-divergences. Almost all of the correlations are above 0.5 and mostly higher. Innovation text sample number 6 is the only exception, its divergences have lower correlation coefficients with all of the other KL-divergences than any other variable.

Table 6 Correlations of the KL-divergences between the topic distributions

	KL 0	KL 1	KL 2	KL 3	KL 4	KL 5	KL 6	KL 7
KL 0	1	0.80	0.90	0.88	0.97	0.79	0.58	0.95
KL 1		1	0.90	0.78	0.79	0.96	0.52	0.84
KL 2			1	0.88	0.87	0.80	0.72	0.88
KL 3				1	0.93	0.75	0.70	0.87
KL 4					1	0.79	0.67	0.90
KL 5						1	0.41	0.84
KL 6							1	0.49
KL 7								1

6.2 Research and Development and Innovation Score

Table 7 presents the correlations between the different text-based innovation measurements and $\frac{RD_{it}}{sales_{it}}$. Innovation measurements 1, 3, 5 and 7 have a correlation coefficient of more than or close to 0.1, which indicates slight correlation. Measurement number 6 is the only one with a negative coefficient.

Table 7 Correlation between Innovation metric and research and development

Correlation Matrix	$\frac{RD_{it}}{sales_{it}}$
$innscore\_0_{it}$	0.062
$innscore\_1_{it}$	0.170
$innscore\_2_{it}$	0.018
$innscore\_3_{it}$	0.089
$innscore\_4_{it}$	0.046
$innscore\_5_{it}$	0.256
$innscore\_6_{it}$	-0.144
$innscore\_7_{it}$	0.215

Table 8 shows regression results for regression function 34, with research and development as a percentage of sales as y-variable. In each column is one regression with a different innovation text-based innovation measurement as $innscore_{it}$. Regressions 1, 2, 3, 5, 6 and 7 show a statistically significant effect of the text-based innovation measurement on the research and development variable at 5% confidence level. However, number 6 has a statistically significant negative effect and so, the ones with the desired positive effect are 1, 2, 3, 5 and 7. In line with the correlation coefficients of table 6, are the results that the innovation score number 2 has a smaller coefficient than the other statistically significant scores. The R-squared and adjusted R-squared are approximately 0.55 for all regressions, which means that the selected variables explain 55% of the variation in the dependent variable.

Table 8 Regression results for R&D

Dependent variable $\frac{RD_{it}}{sales_{it}}$ **observations** 8317

	Regression and innovation score variable number							
	0	1	2	3	4	5	6	7
intercept	20.9476	21.3067	21.2445	21.7642	20.9196	20.9126	20.0631	21.4612
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$innscore_{it}$	0.0036	0.4531	0.1776	0.5634	-0.0232	0.6360	-0.2614	0.6382
	(0.972)	(0.000)	(0.042)	(0.000)	(0.832)	(0.000)	(0.001)	(0.000)
$\frac{P_{it}}{sales_{it}}$	-0.0064	-0.0089	-0.0070	-0.0070	-0.0064	-0.0098	-0.0054	-0.0080
	(0.561)	(0.420)	(0.527)	(0.530)	(0.562)	(0.378)	(0.627)	(0.471)
$log(TA_{it})$	8.6281	8.5058	8.6023	8.568	8.6283	8.4946	8.6268	8.4567
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$log(sales_{it})$	-7.8540	-7.7707	-7.8484	-7.8589	-7.8543	-7.7039	-7.8243	-7.7360
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$log(TL_{it})$	-1.6883	-1.6172	-1.6703	-1.5754	-1.6902	-1.6202	-1.6996	-1.5718
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
ROA_{it}	-0.1948	0.1947	-0.1952	-0.1958	-0.1948	-0.1940	-0.1935	-0.1945
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Industry dummies	X	X	X	X	X	X	X	X
R^2	0.546	0.548	0.546	0.548	0.546	0.549	0.547	0.550
Adjusted R^2	0.545	0.547	0.545	0.547	0.545	0.548	0.546	0.549

The value in the parentheses () represents the p-value of the above coefficient

6.3 Patents and Innovation Score

In Table 9 are the correlation coefficients between the text-based innovation measurement and $\frac{P_{it}}{sales_{it}}$. A positive correlation coefficient of close to or more than 0.1 is between the patent variable and innovation measurement numbers 1, 5 and 7. The correlation coefficient is negative for innovation measurements 3, 4 and 6.

Table 9 Correlation between innovation metric and patents

Correlation matrix	$\frac{P_{it}}{sales_{it}}$
$innscore\_0_{it}$	0.012
$innscore\_1_{it}$	0.124
$innscore\_2_{it}$	0.011
$innscore\_3_{it}$	-0.026
$innscore\_4_{it}$	-0.009
$innscore\_5_{it}$	0.176
$innscore\_6_{it}$	-0.017
$innscore\_7_{it}$	0.082

Regression results for regression function number 3, where patent value as a percentage of sales is the dependent variable, are shown in table 10. The columns represent the 8 different regressions conducted for each innovation text -based innovation measurement. Out of the 8 measurements, numbers 1, 2, 5 and 6 show a statistically significant effect of the innovation measurement on the value of patents and brands at 5% confidence level. The rest of the variables do not have a statistically significant effect.

The R-squared and adjusted R-squared are approximately 0.38 for all regressions of table 10, which means that the independent variables explain about 38% of the changes in the dependent variable. The level is lower for these regressions than the regressions in table 8 with research and development as the dependent variable.

Table 10 Regression results for patents

Dependent variable $\frac{P_{it}}{sales_{it}}$ **observations** 8317

	Regression and innovation score variable number							
	0	1	2	3	4	5	6	7
intercept	13.7830	13.9736	14.0731	13.8470	13.7620	13.7660	14.4206	13.8510
	(0.001)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.001)
$innscore_{it}$	0.0579	0.2576	0.1981	0.0695	0.0246	0.3156	0.2129	0.1066
	(0.573)	(0.001)	(0.022)	(0.458)	(0.821)	(0.000)	(0.008)	(0.139)
$\frac{RD_{it}}{sales_{it}}$	-0.0063	-0.0088	-0.0069	-0.0068	-0.0063	-0.0096	-0.0053	-0.0079
	(0.561)	(0.420)	(0.527)	(0.530)	(0.562)	(0.378)	(0.627)	(0.471)
$log(TA_{it})$	5.0357	4.9825	5.0107	5.0339	5.0379	4.9917	5.0262	5.0214
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$log(sales_{it})$	-7.4219	-7.3858	-7.4177	-7.4287	-7.4241	-7.3635	-7.4344	-7.4150
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$log(TL_{it})$	1.8063	1.8363	1.8209	1.8155	1.8043	1.8279	1.8117	1.8190
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
ROA_{it}	-0.0760	-0.0762	-0.0763	-0.0761	-0.0759	-0.0760	-0.0767	-0.0761
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Industry dummies	X	X	X	X	X	X	X	X
R^2	0.379	0.380	0.380	0.379	0.379	0.380	0.380	0.379
Adjusted R^2	0.378	0.379	0.378	0.378	0.378	0.379	0.379	0.378

The number in the parentheses () represents the p-value of the above coefficient

6.4 Innovation and Performance

Table 11 shows regression results for regression function 7, with the growth of sales in the next year as dependent variable. The innovation measurement shows statistically

significant positive (although very small) results at 95% confidence interval for all regressions. Neither patent nor research and development –variables show statistically significant relationships in any of the regressions. However, this is in line with the results of Bellstam et al. (2019), in their study, the text-based innovation measure predicted firm performance better than patents or research and development rates. The R-squared in these regressions is lower than in the previous ones, only a bit over 3% in all of them. However, there are quite few variables to explain growth of sales, which likely consists of many factors.

Table 11 Firm performance and text-based innovation

Dependent variable $log(sales_{it+1})$ **Observations** *4507*

	Regression and innovation score variable number							
	0	1	2	3	4	5	6	7
intercept	-0.0096	-0.0108	-0.0022	-0.0046	-0.0077	-0.0154	0.0053	-0.009
	(0.924)	(0.916)	(0.982)	(0.964)	(0.939)	(0.879)	(0.958)	(0.929)
$innscore_{it}$	0.0092	0.0059	0.0078	0.0076	0.0082	0.0054	0.0057	0.0078
	(0.002)	(0.006)	(0.002)	(0.004)	(0.008)	(0.025)	(0.015)	(0.000)
$\frac{P_{it}}{sales_{it}}$	-0.0005	-0.0005	-0.0005	-0.0004	-0.0004	-0.0005	-0.0005	-0.0005
	(0.126)	(0.093)	(0.101)	(0.149)	(0.137)	(0.095)	(0.109)	(0.111)
$\frac{RD_{it}}{sales_{it}}$	0.0002	0.0001	0.0002	0.0002	0.0003	0.0002	0.0003	7.07e-5
	(0.457)	(0.651)	(0.544)	(0.564)	(0.426)	(0.640)	(0.374)	(0.830)
$log(sales_{it})$	0.0024	0.0022	0.0017	0.0023	0.0024	0.0025	0.0014	0.0025
	(0.145)	(0.181)	(0.312)	(0.164)	(0.145)	(0.132)	(0.410)	(0.131)
$log(growth_{it-1})$	0.0237	0.0237	0.0239	0.0238	0.0236	0.235	0.0238	0.0237
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Industry dummies	X	X	X	X	X	X	X	X
R^2	0.036	0.036	0.036	0.036	0.036	0.035	0.035	0.037
Adjusted R^2	0.033	0.033	0.033	0.033	0.032	0.032	0.032	0.034

The value in the parentheses () represents the p-value of the above coefficient

6.5 Control Regressions

Table 12 shows regression results for regression functions 5 and 6, respectively, where the Kullback-Leibler divergence from innovation texts is switched to the divergence from a corporate finance text sample multiplied by -1 for easier interpretation. In the first regression with patents as percentage of sales as the dependent variable, a statistically significant connection was not found. In the other regression, with research and development as a percentage of sales as the dependent variable, the relationship is statistically significant, but negative. Thus we can conclude that the method of innovation measurement constructed differs from a measurement based on any corporation related text and should capture actual innovation.

Table 12 Control regressions with corporate finance text

Dependent variable	$\frac{P_{it}}{sales_{it}}$	$\frac{RD_{it}}{sales_{it}}$
intercept	14.4015	16.8163
	(0.000)	(0.000)
$CFscore_{it}$	-0.0068	-0.4593
	(0.938)	(0.000)
$\frac{RD_{it}}{sales_{it}}$	-0.0116	-
	(0.286)	-
$\frac{P_{it}}{sales_{it}}$	-	-0.0118
	-	(0.286)
TA_{it}	5.0539	8.7168
	(0.000)	(0.000)
$sales_{it}$	-7.5587	-8.0455
	(0.000)	(0.000)
TL_{it}	1.9630	-1.5674
	(0.000)	(0.000)
ROA_{it}	-0.0798	-0.1880
	(0.000)	(0.000)
R^2**/adjusted** R^2	0.351/0.351	0.528/0.528

The value in the parentheses () represents the p-value of the above coefficient

6.6 Discussion of Results

Eight different innovation-related text samples were used to test what kind of innovation text works best for measuring innovation. Three different regression models were formed to test the relationship between the text-based innovation measurement and traditional innovation indicators and an increase in sales. Text samples 1, 2 and 5 seemed to form the most robust measurements for innovation based on these results. All of the different measurements have a statistically significant positive effect on the growth of sales, but numbers 1, 2 and 5 were the only ones with a statistically significant positive relationship with both, research and development spending, and the value of patents and brands. The text samples with the worst results were 0, 4 and 6. Numbers 0 and 4 did not have a statistically significant relationship to either one of the traditional innovation indicators and even though number 6 had a positive effect on the patent variable, it had a negative effect on the R&D variable. Number 6 was also the innovation measurement that had lower correlations with the other measurements in table 6.

Text 3 produced an inconclusive result, it was statistically significant on one of the innovation indicator regressions and not significant on one. If we look at the topic distributions in figure 5, we can see that text-based innovation measures 0 and 4, which performed the worst in this study, are the two with the most diverse topic distributions. This could mean that these two texts might have been too general and that they captured too much other aspects than innovation. On the other hand, the distributions of the best-performing measurements 1 and 2 also look quite similar to each other. Texts 6 and 7 have less distributed topic distributions and are clearly constructed of less topics than the other texts, they also produced inconclusive results.

All in all, all of the topic distributions in figure 5 look similar, but even the small differences seem to be significant, judging by the research results. Text samples 4, 6 and 7 were from the different parts of the same innovation textbook and the inconclusive results on all of the text-based innovation measurements from these samples suggests that the book might have been a poor choice.

The positive relationship with the future growth of sales on all of the text-innovation variables could indicate that the measurements that did not produce great correlation results with other innovation indicators, might still not be completely useless. The results leave room for more testing and model development beyond where this study extends. On the other hand, if patents and R&D are not great innovation indicators alone, the correlation with patents and R&D alone should not be what defines a good innovation indicator. As we are trying to construct an indicator that works better than the traditional ones, the validation could concentrate on other factors now that it is proven that the measurement somewhat correlates with these traditional indicators.

There were 8 innovation measurement variables generated and all of them had higher correlations with each other than any of the other innovation indicators. It would still seem more important to have a high correlation amongst the similar innovation measurement, than when compared to other indicators, because they should be expressing the exact same thing. Along with the internal correlation results between the KL-divergences and the fact that measurement number 6 had a statistically significant negative effect on the research and development variable, I would only rule number 6 as unfit for measuring innovation, whereas the other variables showed more promising results.

7 Conclusions

This thesis has studied, whether innovativeness can be measured from the narrative sections of 10-K reports. In the study, the innovation measure developed by Bellstam et al. (2019) was tested with new data. With a modification to their method made in this thesis, the innovation measurement can be made for any firm with a 10-K or annual report with narrative sections. This study extends the literature on innovation indicators and provides a new method of measuring innovation that can be measured for firms that do not hold patents or engage in research and development.

The text-based innovation measurement passed the validation tests for most of the measurements from innovation text samples. The best outcome was that all of the text-based innovation variables predicted sales growth better than patents or R&D. In addition to correlating with other innovation indicators, the ability to predict future sales are important aspects for validating the indicator.

An opportunity for future research on the subject would be to conduct extensive validation for the innovation measurement method. Even though there were industry dummies present in the regressions, more research on industry-related effects could be made. Also, more tests on the relation between the measurement and future innovation, future income and profitability could be made. A thorough validation of an innovation measure is quite difficult though, because the definition of innovativeness itself is still not exhaustive.

There was some variability in the performance of the eight different innovation texts in the regressions. The reason behind this variability should be explored further. The reason could be in the different topic distributions and word frequencies. Some of the texts discuss innovation from different points of view, which seems to affect the results.

Research could also be extended by using a classification method that understands sentiment and for example synonyms. Whereas LDA concentrates on word distributions,

some other method could be used to analyse the context of the words and improve the topic classifications. Bellstam et al. (2019) added a sentiment score to account for the possibility of speaking about innovation in a negative context, and this extension would be useful for this study as well. Based on the topic distributions, some more words could have been removed that do not bring value to the analysis, for example words that appear in every document or only in a few documents.

In conclusion, the study shows promising results on measuring innovation from companies' own disclosure and strategic disclosure or impression management do not seem to significantly distort these results. The research results of this study are encouraging for conducting more research with new data or improved methods.

References

Aggarwal, C. C. & Chai, C. (2012). A Survey of text Classification Algorithms. *Mining Text Data*, 163-222.

Aghion, P., Bechtold, S., Cassar, L. & Herz, H. (2018). The Causal Effects of Competition on Innovation: Experimental Evidence. *The Journal of Law, Economics, and Organization,* 34(2), 162-195.

Aghion, P. & Howitt, P. (1998). *Endogenous Growth Theory.* The MIT Press.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B. & Kochut, K. (2017). *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques.* arXiv: 1707.02919v2. https://arxiv.org/abs/1707.02919

Antonipillai, J., Lee, M. K., Rubinovitz, R., Langdon, D., Yu, F., Hawk, W., Marco, A. C., Toole, A. A. & Tesfayesus, A. (2016). *Intellectual Property and the U.S. Economy: 2016 Update*. Economics & Statistics Administration and U.S. Patent and Trademark Office.

Atkinson, R. D. & Ezell, S. J. (2012). *Innovation Economics: The race for global advantage.* Yale University Press.

Bayarçelik, E. B. & Taşel, F. (2012). Research and Development: Source of Economic Growth. *Procedia - Social and Behavioral Sciences*, 58, 744-753. https://doi.org/10.1016/j.sbspro.2012.09.1052.

Belenzon, S. & Patacconi, A. (2013). Innovation and firm value: An investigation of the changing role of patents, 1985–2007. *Research Policy,* 42(8), 1496-1510. https://doi.org/10.1016/j.respol.2013.05.001.

Belitz, H., Clemens, M., von Hirschhausen, C., Schmidt-Ehmcke, J., Werwatz, A. & Zloczysti, P. (2011). An indicator for national systems of innovation – methodology and application to 17 industrialized countries. *DIW Berlin Discussion Paper*, 1129. http://dx.doi.org/10.2139/ssrn.1858751

Bellstam, G., Bhagat, S. & Cookson, J. A. (2019). A Text-Based Analysis of Corporate Innovation. Working paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2803232

Besimi, A., Dika, Z., Shehu, V., & Selimi, M. (2019). Applied text-mining algorithms for stock price prediction based on financial news articles. *Managing Global Transitions*, 17(4), 335-351, 354-355. http://dx.doi.org.proxy.uwasa.fi/10.26493/1854-6935.17.335-351

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research,* 3, 993-1022.

Buehlmaier, M. M. M., Whited, T. M. (2018). Are Financial Constraints Priced? Evidence from Textual Analysis. *The Review of Financial Studies,* 31(7), 2693–2728, https://doi-org.proxy.uwasa.fi/10.1093/rfs/hhy007

Camilsón, C. & Villar-López, A. (2014). Organizational innovation as an enabler of technological innovation capabilities and firm performance. *Journal of Business Research,* 67(1), 2891-2902.

Chen, Y., Wu, C., Chen, Y., Li, H. & Chen, H. (2017). Enhancement of fraud detection for narratives in annual reports. *International Journal of Accounting Information Systems*, 26, 32-45.

Chiesa, V., Frattini, F., Lazzarotti, V. & Manzini, R. (2009). Performance measurement in R&D: exploring the interplay between measurement objectives, dimensions of performance and contextual factors. *R&D Management*, 39(5), 487–519. https://doi-org.proxy.uwasa.fi/10.1111/j.1467-9310.2009.00554

Cho, C. H., Roberts, R. W. & Patten, D. M. (2010). The language of US corporate environmental disclosure. *Accounting, Organizations and Society,* 35(4), 431-443. https://doi.org/10.1016/j.aos.2009.10.002.

Cohen, L., Diether, K. & Malloy, C. (2013). Misvaluing Innovation. *The Review of Financial Studies*, 26(3), 635-666. www.jstor.org/stable/23355393

Cooper, M. J., Knott, A. M. & Yang, W. (2020). RQ Innovative Efficiency and Firm Value. Working paper. Retrieved 2020-4-17: https://ssrn.com/abstract=2631655

Cortes, C. & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20, 273-297.

Dang, J. & Motohashi, K. (2015). Patent statistics: A good indicator for innovation in China? Patent subsidy program impacts on patent quality. *China Economic Review,* 35, 137-155. https://doi.org/10.1016/j.chieco.2015.03.012.

Diab, D. M. & El hindi, K. M. (2017). Using differential evolution for fine tuning naïve Bayesian classifiers and its application for text classification. *Applied Soft Computing*, 54, 183-199. https://doi.org/10.1016/j.asoc.2016.12.043

Dyer, T., Lang, M., & Stice-Lawrence, L. (2017). The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. Journal of Accounting and Economics, 64(2–3), 221–245. https://doi.org/10.1016/J.JACCECO.2017.07.002

Dziallas, M. & Blind, K. (2019). Innovation indicators throughout the innovation process: An extensive literature analysis. *Technovation*, 81-81, 3-29. https://doi.org/10.1016/j.technovation.2018.05.005

Edison, H., bin Ali, N. & Torkar, R. (2013). Towards innovation measurement in the software industry. *Journal of Systems and Software*, 86(5), 1390-1407. https://doi.org/10.1016/j.jss.2013.01.013.

Enríquez, F., Troyano, J. A. & López-Solaz, T. (2016). An approach to the use of word embeddings in an opinion classification task. *Expert Systems with Applications*, 66, 1-6. https://doi.org/10.1016/j.eswa.2016.09.005

European Commission. (2020). *Community Innovation Survey (CIS).* Eurostat. Retrieved 2020-04-29 https://ec.europa.eu/eurostat/web/microdata/community-innovation-survey

European Patent Office & the European Union Intellectual Property Office (2019). *IPR-intensive industries and economic performance in the European Union. Industry-Level Analysis Report.* 3rd Edition. https://www.epo.org/service-support/publications.html?pubid=201#tab3

Fisher, I. E., Garnsey, M. R. & Hughes, M. E. (2016). Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research. *Intell. Sys. Acc. Fin. Mgmt*., 23, 157– 214.

Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*. Morgan & Claypool.

Gordon, I. R. & McCann, P. (2005). Innovation, agglomeration, and regional development. *Journal of Economic Geography,* 5, 523-543.

Greenhalgh, C. & Rogers, M. (2010). *Innovation, intellectual property, and economic growth.* Princeton University Press.

Griliches, Z. (1990). Patent statistics as economic indicators: A survey. *Journal of Economic Literature*, 28(4), 1661.

Groth, S. S. & Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision support systems*, 50(4), 680-691. https://doi.org/10.1016/j.dss.2010.08.019

Guan, J. & Chen, K. (2010). Measuring the innovation production process: A cross-region empirical study of China's high-tech innovations. *Technovation,* 30(5-6), 348-358. https://doi.org/10.1016/j.technovation.2010.02.001.

Hall, B. H. & Harhoff, D. (2012). Recent Research on the Economics of Patents. *Ann. Rev. Econ.,* 4, 541-565.

Hall, B. H., Helmers, C., Rogers, M. & Sena, V. (2013). The importance (or not) of patents to UK firms. *Oxford Economic Papers,* 65(3), 603 – 629.

Hall, B. H., Jaffe, A. & Trajtenberg, M. (2005). Market value and patent citations. *The Rand Journal of Economics, 36*(1), 16-38.

Heller, M. (2019). Deep learning explained. InfoWorld.Com, Retrieved 2020-05-20 from https://search-proquest-com.proxy.uwasa.fi/docview/2229801791?accountid=14797

Henry, E. (2006). Market Reaction to Verbal Components of Earnings Press Releases: Event Study Using a Predictive Algorithm. *Journal of Emerging Technologies in Accounting,* 3(1), 1-19.

Hoad, T. C. & Zobel, J. (2003). Methods for identifying versioned and plagiarized documents. *Journal of the American Society for Information Science and Technology,* 54, 203-215.

Hoberg, G. & Phillips, G. (2016). Text-Based Network Industries and Endogenous Product Differentiation. Journal of Political Economy, 124(5), p. 1423-1465.

Holland, J. (2009). "Looking behind the veil": Invisible corporate intangibles, stories, structure and the contextual information content of disclosure. *Qualitative Research in Financial Markets*, 1(3).

Hong, S., Oxley, L. & McCann, P. (2012), A Survey of Innovation Surveys. *Journal of Economic Surveys,* 26, p. 420-444.

Hombert, J. & Matray, A. (2018). Can innovation help U.S. manufacturing firms escape import competition from china? *The Journal of Finance,* 73(5), 2003-2039.

Hotho, A., Nürnberger, A. & Paaß, G. (2005). A Brief Survey of Text Mining. Ldv Forum, 20(1), 19-62.

Howitt, P. (2004). Endogenous growth, productivity and economic policy: a progress report. *International productivity monitor,* 8, 3-15.

Huang, H. H., Zang, A. Y. & Zheng, R. (2014). Evidence on the Information Content of Text in Analyst Reports. *The Accounting Review*, 89(6), 2151-2180. https://doi.org/10.2308/accr-50833

Humphreys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K. & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems*, 50(3), 585-594. https://doi.org/10.1016/j.dss.2010.08.009

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2017). *An Introduction to Statistical Learning with Applications in R.* Springer.

Janssen, M.C. and Roy, S. (2015). Competition, Disclosure and Signalling. *The Economic Journal,* 125, 86-114.

Junge, M., Severgnini, B. & Sørensen, A. (2016). Product-Marketing Innovation, Skills, and Firm Productivity Growth. *Review of Income and Wealth,* 62, 724-757.

Kamaruddin, S. S., Abu Bakar, A., Hamdan, A. R., Nor, F. M., Nazri, M. Z. A., Othman, Z. A. & Hussein, G. S. (2015). A text mining system for deviation detection in financial documents. *Intelligent Data Analysis,* 19, S19–S44. https://doi-org.proxy.uwasa.fi/10.3233/IDA-150768

Kim, S. & Kim D. (2014). Investor sentiment from internet message postings and the predictability of stock returns. *Journal of Economic Behavior & Organization*, 107(B), 708-729. https://doi.org/10.1016/j.jebo.2014.04.015

Kirk, M. (2017). *Thoughtful Machine Learning with Python: A Test Driven Approach.* O'Reilly.

Kogan, L., Papanikolaou, D., Seru, A., & Stoffman, N. (2017). Technological Innovation, Resource Allocation, and Growth. *Quarterly Journal of Economics*, 132(2), 665–712. https://doi-org.proxy.uwasa.fi/10.1093/qje/qjw040

Kuznets, S. (1969). *Modern Economic Growth: Rate, Structure and Spread*. Yale University Pres.

Le, Q. & Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on International Conference on Machine Learning, 32,* II–1188–II–1196

Leung, S., Parker, L. & Courtis, J. (2015). Impression management through minimal narrative disclosure in annual reports. *The British Accounting Review,* 47, 275-289. https://doi.org/10.1016/j.bar.2015.04.002

Lewis, C. & Young, S. (2019). Fad or future? Automated analysis of financial text and its implications for corporate reporting. *Accounting and Business Research,* 49(5), 587-615. https://doi.org/10.1080/00014788.2019.1611730

Lobo, G.J., Xie, Y. & Zhang, J.H. (2018). Innovation, financial reporting quality, and audit quality. *Review of Quantitative Finance and Accounting,* 51, 719–749. https://doi-org.proxy.uwasa.fi/10.1007/s11156-017-0686-1

Lopez-Gazpio, I., Maritxalar, M., Gonzalez-Agirre, A., Rigau, G., Uria, L. & Agirre, E. (2017). Interpretable semantic textual similarity: Finding and explaining differences between sentences. *Knowledge-Based Systems,* 119, 186-199. https://doi.org/10.1016/j.knosys.2016.12.013

Loughran, T. & McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, 54, 1187-1230. https://doi-org.proxy.uwasa.fi/10.1111/1475-679X.12123

Lowry, M., Michaely, R. & Volkova, E. (2019). Information Revelation through Regulatory Process: Interactions Between the SEC and Companies Ahead of the IPO. Fifth

Annual Conference on Financial Market Regulation, Swiss Finance Institute Research Paper 19-47. http://dx.doi.org/10.2139/ssrn.2802599

Majumder, G., Pakray, P., Gelbukh, A. & Pinto, D. (2016). Semantic Textual Similarity Methods, Tools, and Applications: A Survey. *Computación y Sistemas,* 20(4), 647-665.

Matin, R., Hansen, C., Hansen, C. & Mølgaard, P. (2019). Predicting distresses using deep learning of text segments in annual reports. *Expert Systems with Applications*, 132, 199-208. https://doi.org/10.1016/j.eswa.2019.04.071

Merkl-Davies, D., & Brennan, N. M. (2007). Discretionary disclosure strategies in corporate narratives: Incremental information or impression management? Journal of Accounting Literature, 26, 116-194.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems, vol. 2,* Red Hook, NY (3111-3119). Curran Associates Inc.

Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications,* 106, 36–54. https://doi.org/10.1016/j.eswa.2018.03.058

Moratis, L. (2018). Signalling responsibility? Applying signalling theory to the ISO 26000 standard for social responsibility. *Sustainability,* 10(11) doi:http://dx.doi.org.proxy.uwasa.fi/10.3390/su10114172

Mukherjee, A., Singh, M. & Žaldokas, A. (2017). Do corporate taxes hinder innovation?. *Journal of Financial Economics*, 124(1), 195-221. https://doi.org/10.1016/j.jfineco.2017.01.004.

Nagaoka, S., Motohashi, K., Goto, A. (2010). Patent Statistics as an Innovation Indicator. *Handbook of the Economics of Innovation,* 2, 1083-1127. https://doi.org/10.1016/S0169-7218(10)02009-5

Nakache D., Metais E., Timsit J.F. (2005) Evaluation and NLP. In: Andersen K.V., Debenham J., Wagner R. (eds.), *Database and Expert Systems Applications. DEXA 2005. Lecture Notes in Computer Science,* vol. 3588 (p. 626-632). Springer.

OECD. (2005). *INNOVATION.* Glossary of statistical terms. Retrieved 2020-04-17 https://stats.oecd.org/glossary/detail.asp?ID=6865

Onan, A., Korukoglu, S. & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, 57, 232-247.

Park, K., Seok Hong, J. & Kim, W. (2020) A Methodology Combining Cosine Similarity with Classifier for Text Classification. *Applied Artificial Intelligence,* 34(5), 396-41.

Patelli, L. & Pedrini, M. (2014). Is the Optimism in CEO's Letters to Shareholders Sincere? Impression Management versus Communicative Action during the Economic Crisis. *Journal of business ethics*, 124, 19-34.

Peterson, K., Schmardebeck, R. & Wilks, T. J. (2015). The Earnings Quality and Information Processing Effects of Accounting Consistency. *Accounting Review*, *90*(6), 2483–2514. https://doi-org.proxy.uwasa.fi/10.2308/accr-51048

PRH. (2019). *What kind of inventions can be patented?* Finnish patent and registration office. Retrieved 2020-05-29 https://www.prh.fi/en/patentit/theabcofpatenting/whatcanbepatented.html.

Purda, L. and Skillicorn, D. (2015). Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection. *Contemporary Accounting Research,* 32, 1193-1223.

Refinitiv Eikon. (2018). [*Financial variables: Net sales or revenues; Net sales/revenunes growth; Return on assets; Research & Development/Sales; Brands , Patents – Gross; Total liabilities; Total assets*]. Retrieved 2020-08-03 from https://eikon.thomsonreuters.com/index.html [Requires for the database to be installed]

Roper, S. & Hewitt-Dundas, N. (2015). Knowledge stocks, knowledge flows and innovation: Evidence from matched patents and innovation panel data. *Research Policy,* 44(7), 1327-1340. https://doi.org/10.1016/j.respol.2015.03.003.

Rönnqvist, S. & Sarlin, P. (2017). Bank distress in the news: Describing events through deep learning. *Neurocomputing,* 264, 57-70.

Schumpeter, J. A. (1943). *Capitalism, Socialism and Democracy.* Harper & Brothers.

Sokolova, M. & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing and management,* 45, 427-437.

Statistics Finland (2020). *Innovation survey: Questionnaires.* Retrieved 2020-09-16. https://www.stat.fi/keruu/inno/lomakkeet_en.html

Taques, F. H., López, M. G., Basso, L. F., Areal, N. (2020). Indicators used to measure service innovation and manufacturing innovation. *Journal of Innovation &*

Knowledge, advance online publication. https://doi.org/10.1016/j.jik.2019.12.001

Thompson, V., Panchev, C. & Oakes, M. (2015). Performance Evaluation of Similarity Measures on Similar and Dissimilar Text Retrieval. *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR,* Lisbon, Portugal (p. 577-584).

Tsai, M. & Wang, C. (2017). On the risk prediction and analysis of soft information in finance reports. *European Journal of Operational Research,* 257, 243-250. https://doi.org/10.1016/j.ejor.2016.06.069

SICCODE. (2020). *Structure of SIC CODE 35 – Industrial Machinery and Equipment*. Retrieved 2020-07-13 https://siccode.com/sic-code-hierarchy/35/industrial-commercial-machinery-computer-equipment

U.S. Securities and Exchange Commission. (2011). *How to Read a 10-K.* Retrieved 2020-11-06 https://www.investor.gov/introduction-investing/general-resources/news-alerts/alerts-bulletins/investor-bulletins/how-read

U.S. Securities and Exchange Commission. (2018). *Using EDGAR to Research Investments.* Retrieved 2020-06-11 https://www.sec.gov/oiea/Article/edgarguide.html

U.S. Securities and Exchange Commission. (2020). *[10-K filings].* Retrieved 2020-08-03 from https://www.sec.gov/edgar.shtml

USPTO. (2015). *General Information Concerning patents*. United States Patent and Trademark Office. https://www.uspto.gov/patents-getting-started/general-information-concerning-patents

Wang, T., Ulmer, J. R. & Kannan, K. (2013) The Textual Contents of Media Reports of Information Security Breaches and Profitable Short-Term Investment Opportunities. *Journal of Organizational Computing and Electronic Commerce,* 23(3), 200-223. https://doi-org.proxy.uwasa.fi/10.1080/03610918.2013.833227

Weiss, P. (2003). Adoption of product and process innovations in differentiated markets: The impact of competition. *Review of Industrial Organization, 23*(3-4), 301-314.

WIPO. (2020). *Guide to the international patent classification.* World Intellectual Property Organization.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*. Elsevier Science & Technology.

Wong, T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern recognition,* 48(9), 2839-2846.

Xu, S. (2018). Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*, 44(1), 48-59.

Yang, Y. & Chute, C. G. (1994). An example-based mapping method for text categorization and retrieval. *ACM Transactions on Information Systems,* 12(3), 252-277. https://doi-org.proxy.uwasa.fi/10.1145/183422.183424

Zhu, H., Zhao, S. & Abbas, A. (2019). Relationship between R&D grants, R&D investment, and innovation performance: The moderating effect of absorptive capacity. Journal of Public Affairs, 20.

Appendix

List of the innovation text samples

Text no.	Source	Chapters in text sample
0	Link, A. N. and Siegel, D. S. (2007). *Innovation, Entrepreneurship, and Technological Change*. Oxford University Press Inc.	Chapters 1-3.
1	Talukder, M. (2014). *Managing Innovation Adoption: From Innovation to Implementation.* Taylor & Francis Group.	Chapter 1.
2	Tidd, J. (2015). *Innovation and Entrepreneurship*. Wiley Textbooks.	Chapter 3.
3	Machado, C. & Davim J. P. (2015). *Innovation Management: In Research and Industry.* De Gruyter, Inc.	Chapter 1.
4	Atkinson, R. D. and Ezell, S. J. (2012). *Innovation Economics: The Race for Global Advantage.* Yale University Press.	Chapter 9.
5	Korres, G. M. (2012). *Handbook of Innovation Economics.* Nova Science Publishers, Inc.	Chapter 1.
6	Atkinson, R. D. and Ezell, S. J. (2012). *Innovation Economics: The Race for Global Advantage.* Yale University Press.	Chapter 1.
7	Atkinson, R. D. and Ezell, S. J. (2012). *Innovation Economics: The Race for Global Advantage.* Yale University Press.	Chapter 6.