Università
degli Studi
della Campania
Luigi Vanvitelli

**Dipartimento
di Economia**

PH.D. IN
ENTREPRENEURSHIP AND INNOVATION
XXXVth CYCLE

Area 13 – Scienze Economiche e Statistiche
SECS – P/07

Doctoral Thesis in

INFORMETRICS FOR FINANCIAL ANALYSIS: EXTRACT KNOWLEDGE FROM NARRATIVE AND NUMBERS

Coordinator
Prof. Marcello Martinez

Tutor
Prof. Corrado Cuccurullo

Candidate
Dott.ssa. Alessandra Belfiore

A.Y.
2022/2023

Table of Content

1. Introduction

1.1 Thesis general framework

1.2 Thesis structure

1.2.1 Text mining in accounting and finance

1.2.2 Financial configurations analysis

1.2.3 Narrative Analysis

1.3 Originality

2. A bibliometric review of text mining in accounting and financial research

2.1 Introduction

2.2 Theoretical background

2.2.1 Text mining

2.2.2 Conceptual Framework

2.3 Methods

2.3.1 Data Collection

2.3.2 Data Analysis

2.4 Findings

2.4.1 Descriptive findings

2.4.2 Conceptual structure: thematic evolution

2.4.2.1 (A) A&F/High Impact

2.4.2.2 (B) A&F/Low Impact

2.4.2.3 (C) Management/High Impact

2.4.2.4 (D) Management/Low Impact

2.5 Conclusions

2.6 References

3. Financial Configurations of Italian private hospitals: an evolutionary analysis

3.1 Introduction

3.2 Background

3.2.1. Current knowledge about private hospitals

3.2.2 Towards a practical framework

3.2.2.1 System-wide factors in Italy

3.2.2.2 Hospital-level factors

3.3 Materials and Methods

3.3.1 Data Collection

3.3.2 Data Analysis

3.3.2.1 Partial Triadic Analysis (PTA)

3.3.2.2 Hierarchical clustering

3.3.2.3 Configuration evolution

3.4 Findings

3.4.1 Ownership structure of private hospitals

3.4.2 Discovering main financial dimensions

3.4.3 Main financial configurations

3.4.4 Tracing configurations' paths

3.5 Discussion

3.6 Conclusions

3.7 References

3.8 Annex

3.8.1 Elsevier Copyright: Free use and share for scholarly purposes, included a thesis.

3.8.2 RightsLink by Copyright Clearance Center associated to the paper.

3.8.3 Permission of use released by co-authors.

4. Emotional Language and Financial Performance: Evidence from CEO Discourse during the Covid-19 Pandemic

4.1 Introduction

4.2 Literature Review

4.3 Methods

4.3.1 Data Collection

4.3.2 Data Analysis

4.3.2.1 Topic Modeling

4.3.2.2 Emotional analysis and Sentiment analysis

4.4 Findings

4.4.1 Descriptive findings

4.4.2 Topic context

4.4.3 Relationship between emotional/sentimental language and performance

4.5 Conclusions

4.6 References

5. Conclusions

5.1 Discussion

5.2 Theoretical and Practical implications

5.3 Future research activities

CHAPTER 1

Introduction

Abstract

Financial analysis is evolving with digital technology, enabling organizations to make more informed decisions. The use of big data, unstructured data and automated data analysis are transforming financial analysis, making it more efficient and effective. This thesis focuses on informetrics in accounting and finance, which uses data science and accounting and financial knowledge to measure information and transform it into knowledge. The interdisciplinary approach outlined in the thesis can lead to a more comprehensive and holistic understanding of financial statements analysis, reducing the risk of bias and subjectivity. The thesis offers also an innovative approach to financial statements analysis by proposing a TALL-FinStat (Text Analysis for All - Financial Statement) application that will allow anyone to conduct textual analysis of financial data. Future research should focus on integrating unstructured and structured data, developing ad hoc techniques, and ensuring transparency in the use of advanced technologies in financial analysis. In summary, this work contributes to the field of financial statements analysis and it is the basis for future research in this area.

Keywords: Informetrics - Text analysis - Financial analysis - Unstructured data

1. Introduction

1.1 Thesis general framework

Financial analysis has been a discipline that has developed over time, evolving with technological innovation and the availability of new tools and methods of analysis. In the past, financial analysis was primarily based on traditional financial data, which were extracted from organizations' financial statements. This data was then analyzed through quantitative methods such as ratios and margins, which helped assess the performance and trends of organizations and the market. However, this process was often tedious and ineffective, as data were collected manually from sources such as financial books and records, which required a lot of resources and time. Financial analysts had fewer obligations for transparency and communication to investors and corporate decision makers, which made financial analysis untruthful. Analysts also focused mainly on traditional financial data, neglecting investor opinions, industry trends, and general economic conditions. This was due to the lack of tools to collect, store, and analyze other types of information. Financial analysis is evolving to meet the needs of business decision makers and investors to analyze all available data more efficiently and to account for changes in the global economic environment.

The evolution of financial analysis is driven by:

a) Big Data

The evolution of financial analysis is being driven by the availability of large volumes of business information, so-called "big data". The growth of digital technologies is enabling organizations to collect and use an increasing amount of data, from transaction information to customer opinions on social media. The availability of big data offers the opportunity for organizations to use an unprecedented amount of information thus enabling them to make more informed decisions. In fact, big data makes it possible to tailor analysis to a organization’s specific needs, providing more relevant and useful information. This allows business and financial analysts to focus on the most important data and quickly identify opportunities and risks to communicate to the company. By analyzing data in real time, analysts empower organizations to react quickly to market shocks and make more timely decisions. In addition, big data also enables predictive analytics, which uses historical data and machine learning algorithms to predict future events. Finally, financial analysis based

on big data can be used to help organizations develop new products and services, enabling organizations to remain competitive and grow in an increasingly complex financial environment.

b) Unstructured Data

Financial analysis is increasingly moving away from the classical view focused on the analysis of financial statements ratios alone. The focus on information from accounting notes, auditors' reports, and management letters has led organizations to employ a wide range of tools and technologies to be able to analyze so-called unstructured data. Tools including Natural Language Processing (NLP) algorithms and sentiment analysis enable them to process this data more efficiently and identify opportunities that may not be apparent by analyzing traditional numerical data. Analyzing unstructured data, such as text, allows for greater flexibility in the investigation of financial data because unstructured data are less constrained by the restrictions of structured data. Thus, analysts can explore new data sources and use a variety of analysis methods to obtain additional information about the company and help identify potential trends and risks.

c) Automated data analytics

In financial analysis, one of the biggest challenges in using big data and unstructured data is their collection, management, and interpretation. The amount of information available can be overwhelming and difficult to manage, and financial analysts often need advanced analytical tools to combine and analyze the data effectively. It is important to consider the quality of the data used and verify its reliability before basing decisions on it. With the advent of digitization and electronic data transmission, access to financial information has become increasingly easy, facilitating data collection and analysis. Organizations and investors are increasingly relying on automated data analysis to make strategic decisions, recognizing the importance of having access to reliable and timely information to support their businesses. Automated data analytics also has the potential to reduce errors and wait times in data collection and processing. Ultimately, the use of advanced technologies, such as data science and machine learning, is transforming financial analysis, making it increasingly accessible, efficient, and effective for organizations and investors.

1.2 Thesis structure

To meet the new demands arising from automated data analysis and access to a wide and diverse range of data, financial analysis must necessarily use an interdisciplinary domain (Figure 1). This domain has emerged as an interdisciplinary area that follows an information science workflow, uses data science methods whose knowledge is provided by computer science tools, in a specific domain which is that of accounting and finance. This interdisciplinary area takes the name of informetrics in accounting and finance, or the ability to measure information to transform it into knowledge in accounting and finance. Informetrics in accounting and finance uses innovative research methods to collect, process and analyze large amounts of data, computer software to be able to use different types of data (structured and unstructured data) and accounting and financial knowledge to be able to interpret the results. The entire process follows a classic information science workflow that from the collection arrives at the synthesis of the results. Specifically, this field includes various data science methods, such as statistical analysis, machine learning, and network analysis, to extract insights and models from large financial datasets. Computer tools such as data mining and natural language processing are used instead to analyze and extract relevant information from financial reports, news articles and other financial documents. Using a combination of quantitative methods and data-driven insights, financial data analysts can provide more truthful insight into organizations and the market.

Informetrics in accounting and finance represents a step forward in financial analysis. Tailoring analytics to specific business needs, real-time analytics, and predictive analytics are just some of the benefits of this domain. Informetrics in accounting and finance is, therefore, an essential domain for organizations that want to take full advantage of the opportunities offered by digital technology and big data to improve their financial performance and decision-making capabilities.

Given the informetrics framework I have illustrated, the whole thesis work is developed as a multi-year research project on quantitative methods applied to data, structured and unstructured, extracted from companies' financial records. While the scientific literature on quantitative analysis of financial statements is very rich, still little has been explored about the infinite potential of methodological applications of data science on textual data in accounting and finance.

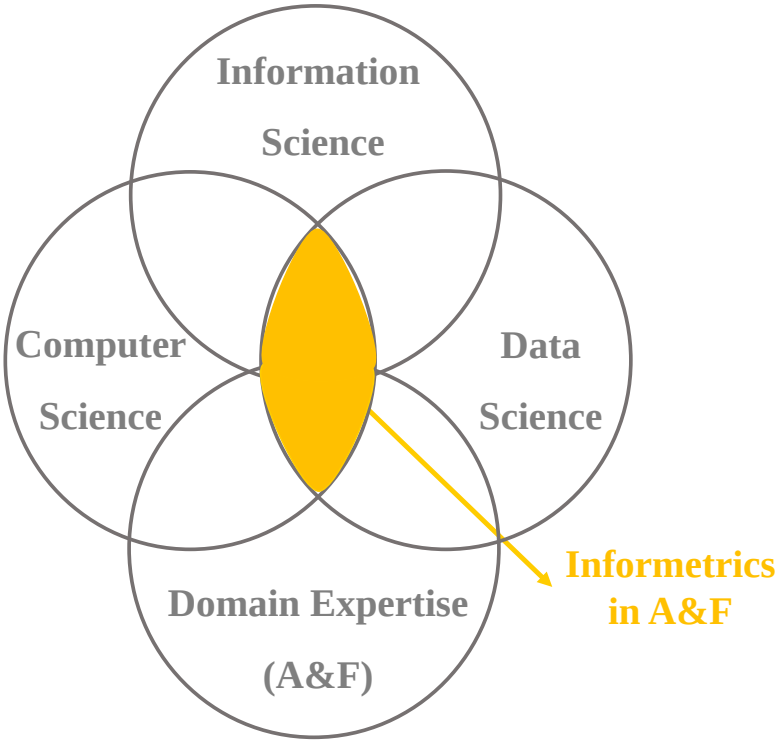


Figure 1 Informetrics in Accounting and Finance

1.2.1 Text mining in accounting and finance

The second chapter conducts a bibliometric analysis with the aim of assessing the scientific productivity of articles on text mining in accounting and finance, during the period 2011-2022. I use sources (i.e., journals) and subject categories (accounting and finance vs. management) to focus the analysis. This research aims to examine the diverse knowledge published in journals, with different levels of impact, to identify key issues that promote the use of text mining methods in accounting and finance vs management disciplines. Using the analysis of the thematic evolution of publications for the four subsections considered, I analyze the motor themes, central themes, niche themes, and emerging or declining themes. Building on the two seminal contributions by Loughran and McDonald (2011; 2016), I compare the thematic map over two periods, analyze and map the evolution of themes (in terms of trajectory over time). Several key insights can be drawn from this study.

1) The topic is extremely recent. A burst in the literature production starts in 2019.
2) The topic is particularly flourishing in the subset of publications in the management category in high-impact journals.

3) The analysis of the thematic evolution of the publications, for the two sub-periods considered, identified several main themes between high-impact and low-impact accounting and finance and management publications.
   3.1 High-impact accounting and finance publications cover corporate social responsibility, non-financial reporting, environmental disclosure, while low-impact publications cover earnings, accounting information, CEO overconfidence, decision-making, and business risk.
   3.2 Publications in the high-impact management sources cover bibliometrics, riviews and, online, language and communication, and textual analysis, while low-impact ones cover consumer knowledge, word of mouth, and networks.

This paper, which discusses the theoretical context of text mining with a framework on methodologies used in accounting and finance, was the basis for designing and developing of TALL-FinStat (Text Analysis for All - Financial Statement). It is an R package for text analysis of financial statements. The goal is to make an application, which is easy to use even for those who do not have programming skills in R. This application will make it possible to analyze, with quantitative applications, numerical and textual data in the field of accounting and finance, following the workflow proposed in the master thesis work (Figure 2).

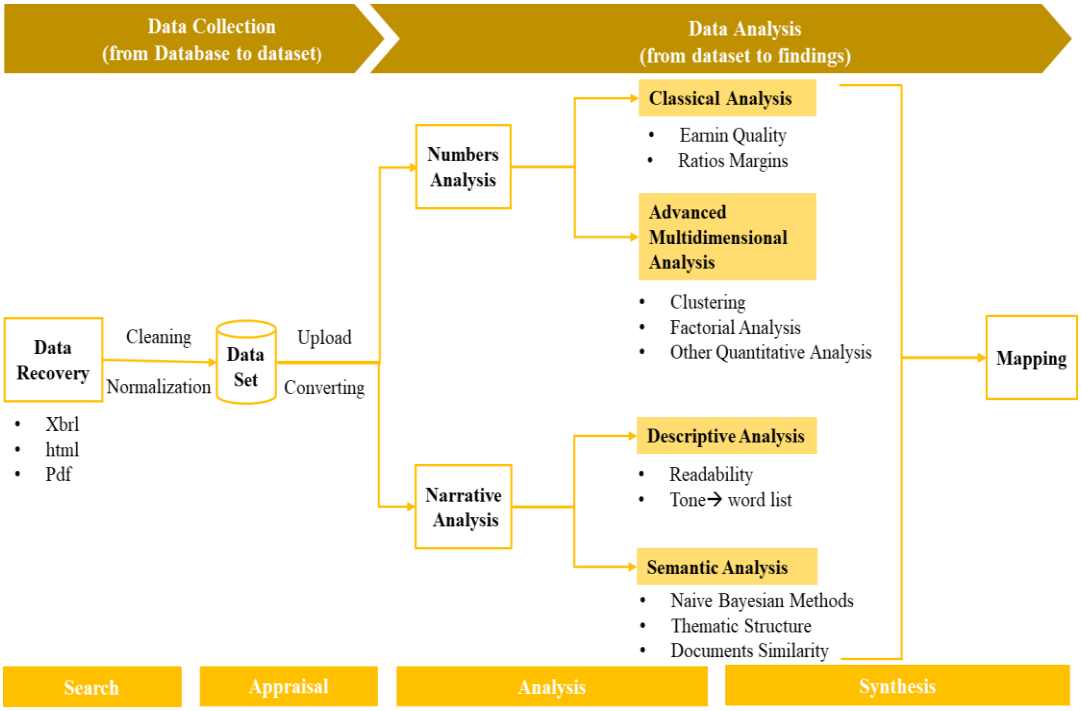


Figure 2 TALL-FinStat Workflow

1.2.2 Financial configurations analysis

The third chapter combines partial triadic analysis (PTA) and cluster analysis, using a configurational approach on the financial data of small and medium-sized enterprises in the Italian accredited private hospital sector. The goal is to show how the use of data science methods in accounting and finance can be useful for pairing hospitals with the same financial configurations and then studying their evolution over time. The study makes some relevant contributions.

The study shows:

1) Four main financial configurations are identified, three of which worsened during the macroeconomic crisis of 2008-2016.
2) It shows that Italian private hospitals have unsustainable business models and are highly exposed to banks, so they need to retrain to benefit from operational and financial synergies.
3) It highlights the need to increase the professionalism of the management structure of private hospitals through the opening of share capital to non-family shareholders.

4) The paper suggests combining clustering analysis with qualitative comparative analysis to find more interesting and complex configurations, providing potential for future research.

This paper, which demonstrates the usefulness of applying data science to accounting and finance data, is one of the academic publications discussed in the Data Science in Accounting and Finance Lab (DAF Lab). The lab will focus on some quantitative research methodologies. It explores traditional and non-traditional techniques and tools in the field of accounting and finance. Through the discussion of different academic research, an innovative workflow for the preparation visualization and analysis of digital financial data and specific data science techniques are presented.

1.2.3 Narrative analysis

The fourth chapter aims to identify the communication strategies adopted by companies during the Covid-19 pandemic. The analysis is based on letters to shareholders contained in the annual reports of companies listed in the Standard & Poor's 500, Euro Stoxx and Nikkei indexes for the year 2020. Topic modeling, emotional analysis along with sentiment analysis and correlation analysis with Total Return provide a comprehensive view of the strategies and themes adopted by companies in response to the pandemic.

The results of the study show that:

1) The 998 companies in the sample were classified into 12 sectors. Prominent among them is the prevalence of the industrial and financial sectors.

2) In relation to performance, I found that U.S. companies more than others suffered the negative effects of the pandemic.

3) Textual analysis showed that CEO letters were rich in information but written clearly and concisely.

4) Topic modeling highlighted 14 significant topics. The topics range from corporate governance and economic performance to sustainability, digital transformation, and customer relations.

5) Analysis of emotions and sentiments, and the correlation between them, suggest that CEOs focused on building trust and conveying a sense of hope and optimism for the future while facing significant challenges.

This initial inquiry presents the basis for two forthcoming research directions. An initial research idea for the future is to broaden the time window of the analysis. Probably the effect of the Covid-19 pandemic will have changed these communication strategies even in the years after 2020. A second line of research is to investigate other characteristics of these companies. For example, using structural, operational, and financial information, it will be possible through a configurational study to identify different communication strategies. The conceptual and practical implications of this work are significant. The study suggests that analysis of financial statements should go beyond conventional ratios and margins to include textual analysis of financial documents to identify hidden patterns and trends. The study also recommends that policymakers encourage and finance industry restructuring, targeting digitization and sustainability, to enable companies to benefit from long-term growth and success. Companies should adapt their investment strategies, including investing in digital technologies and remote work platforms to ensure business continuity in the event of future crises or disruptions. Finally, companies should maintain a strong customer orientation and adapt to their changing needs.

1.3 Originality

The work originality is twofold. It consists in having examined an interdisciplinary flourishing domain, and in having outlined a workflow for the text analysis of financial reports. By adopting an interdisciplinary perspective, the study explores the domain of accounting and finance informatrics. The interdisciplinary perspective of the study therefore allows for a more complete and holistic understanding of financial report analysis, which can lead to better financial analyses. Furthermore, the workflow offered in this research will allow the construction of an application for quantitative research, integrating both textual and numerical analytical methods, which could further improve financial statements analyses.

This quantitative research application provides a more objective, data-driven approach to analyzing financial statements, which can help reduce the risk of bias and subjectivity in financial analysis.

Overall, the study offers a valuable contribution to the field of financial report analysis, providing a new and innovative approach to use informetrics in accounting and

finance. The interdisciplinary perspective, systematic protocol, and application of quantitative research presented in the study can help improve the quality and accuracy of financial analysis and open the way for future research in this area.

CHAPTER 2

A bibliometric review of text mining in accounting and financial research

Abstract

Text mining has gained an increasingly important role in accounting and financial information. Research on text mining in accounting and finance has attracted much attention given the considerable literature. This study aims to systematically review the results of existing academic research in the field, using bibliometric techniques. Specifically, provides a focus on the domain and monitoring of thematic development research. This study contributes in several ways to the field of improved financial and accounting research in text mining. First, it provides the latest research status for researchers interested in the field through the literature review. Second, it helps scholars become more aware of research topics by identifying different topics for different levels of source impact. Third, it provides insights for researchers engaged in the field and motivates collaboration with other disciplines.

Keywords: Text mining - Accounting and Finance - Source Impact - Bibliometric analysis - Interdisciplinary Research

2.1 Introduction

A large amount of textual data is generated every day in accounting and finance. An example is the obvious increase in the length of analyst reports, shareholder letters, website contents and social media posts by corporate stakeholders (Lewis and Young, 2019). Most of this text data is unstructured and spread across the web (Shah et al. 2020, Salloum et al. 2017, Kinsella et al. 2011). This availability of data creates opportunities and challenges for both practitioners and scholars in the fild. Indeed, if this textual information is collected, structured, processed and analysed correctly, valuable knowledge can be gained (Jeseena and David, 2014).

Text mining combines statistical, computer science and linguistic methods to analyse large volumes of text and obtain patterns to be evaluated and interpreted in intelligent applications (Talib et al. 2016). In recent years, text mining applications have multiplied, across all disciplines. Its great potential of text mining is mainly exploited by big companies, academics and governments that want to use the enormous hidden knowledge to make crucial decisions. These subjects prefer automated approaches to analysing their data and the latent information it contains. ML and DL algorithms are just some of the many algorithms that can be used.

In the accounting and finance literature, text mining applications have been carried out in a non-integrated manner, relying on simple methodologies to support results from quantitative applications. Examples include: the evaluation of analysts' forecasts (Muslu et al., 2019), future financial performance (Clarkson et al., 2019; Manchiraju and Rajgopal, 2017) and contemporary company performance (Wang et al., 2018; Nazari et al., 2017), corporate disclosures (Bauguess, 2018; Hoberg and Lewis, 2017) and even labour productivity (Christensen et al., 2017). Text mining, used to obtain qualitative insights from unstructured texts, and text analysis, used to obtain quantitative results, are thus emerging as complementary investigative strategies to the analysis of financial statements metrics, for the study of corporate data.

However, as the use of these textual data analysis techniques becomes more widespread, much needs to be done to organize research on the topic. Bibliometric analysis is an effective and widely applied tool for examining the state of the art. Therefore, this study conducts a bibliometric analysis with the aim of assessing the state of the art of text mining research in accounting and finance from the two seminal

contributions by Loughran and McDonald (2011; 2016). This research aims to examine the diverse knowledge published in multiple journals in recent years to identify key issues that promote the use of text mining techniques in accounting and finance and other related disciplines. The main objectives of this study include (1) a domain focus and (2) monitoring research on thematic development.

This study is structured as follows. Section 2 presents the theoretical background. Section 3 presents the materials and research methods I used to collect and analyse the data. Section 4 describes the main findings and section 5 discusses them with implications for research.

2.2 Theoretical background

2.2.1 Text mining

Text mining methods can be considered an extension of data mining to text data (Romero and Ventura, 2007). It is a technique used for the discovery of hidden information in texts (Hearst, 2003). The main objective of text mining is to analyse large volumes of text and identify patterns for the evaluation and interpretation of the new information found. Thus, through computational power of computers, which adopt statistical and data management algorithms, text mining enables researchers to extract meaning and classify topics more quickly. The tools that can be used vary from those that are fully automated to those where the knowledge of professionals helps in the process of interpreting textual big data.

Within text mining fall the categories of (i) targeted phrases, (ii) sentiment analysis, (iii) topic modelling, (iv) document similarity measures, and (v) readability.

Targeted phrases, also known as specific phrase mining, is a text mining technique of extracting phrases that contain relevant information from a text (Gambhir, Gupta, 2017; Allahyari et al., 2017). These phrases can be used to summarise the content of the text, identify important information or for other purposes. The process of extracting targeted sentences can be performed using various techniques, including the use of clustering algorithms, semantic analysis and the use of machine learning models.

Sentiment analysis is a text mining technique that involves analysing the emotion or attitude expressed in a text to determine whether it is positive, negative or neutral (Liu, 2020; Liu, 2012; Liu, Zhang, 2012). This technique can be used to analyse web pages,

customer reviews, letters to shareholders and other types of text to understand users' opinions and perceptions. Sentiment analysis can be performed using various techniques, including the use of classification algorithms, semantic analysis and the use of machine learning models.

Topic modelling is a text mining technique that consists of identifying the main themes present in a set of documents (Mohr, Bogdanov, 2013; Blei, 2012; Dou et al., 2011). This technique uses natural language processing algorithms to analyse the textual content and group documents according to themes. Topic modelling can be used to analyse large amounts of textual data, such as newspaper articles, books, annual reports and more.

Document similarity measurement is a text mining technique that consists of using algorithms to determine the similarity between two or more documents (Gomaa, Fahmy, 2013; Huang, 2008; Manning; 2008). This technique can be used to group similar documents, to find duplicates or for other purposes. Document similarity measures can be used to compare texts written in different languages or to compare texts of different genres. There are several techniques for calculating similarity between documents, including content comparison based on keywords, comparison based on semantic content and comparison based on text structure.

Readability is a text mining technique that consists of using algorithms to assess the readability of a text. Readability is a measure of the comprehensibility of a text and can be used to assess the accessibility of a text for a particular audience or to identify texts that might be made deliberately difficult to understand (Crossley et al., 2017; Kincaid et al., 1975; Senter, Smith, 1967). There are several methods to calculate readability, including the use of mathematical formulas based on word and sentence length, the use of machine learning models and the use of comprehension tests.

These techniques have great potential to improve know-how and advance entire fields through the processing of large amounts of text data.

2.2.2 Framework

Text mining is already a reality in many fields but much needs to be done to fully exploit its potential in accounting and finance. The question was posed by Loughran and McDonald (2011) when, analysing the tools traditionally used to assess managerial

tone in business documents, they discovered that most of the words that form the lists used to measure sentiment in texts, do not have the same meaning when used in the context of accounting and financial documents. They report that many of the negative words may not be, or may even be proxies for specific sectors, in accounting and finance. In fact, they propose the first lists of words created by examining the use of words in a sample of annual documents relating to companies listed on the US stock market. They thus create sentiment lists based on the most likely interpretation of a word in a corporate context. Their results aim to emphasise that textual analysis can contribute to the ability of accounting and financial professionals to understand hidden information in texts or otherwise extract incremental information. This led the authors in 2016 to conduct a literature review in which they describe in depth the textual analysis methods that characterise existing research on corporate finance (Loughran and McDonald, 2016). Again, they show how the Fog index, generally used to measure the readability of texts, is ineffective for business documents. Their results underline the need for business theory to guide the choice of methods. In particular, the inaccuracy of textual analysis should not preclude its use in accounting and finance. Starting from the comparison with the tools used in other disciplines, specific tools must be created to achieve empirical results that have a credible impact and can be reasonably replicated.

Since these two studies, which represent milestones for accounting and finance researchers, the applications of text mining have multiplied. Accounting and finance researchers have used text mining to analyse: the relationship between the thematic content of financial statements disclosures and predictions of financial errors (Brown et al., 2020; Guay et al., 2016); the relationship between the complexity of corporate social responsibility (CSR) reports and performance (Du and Yu, 2021; Clarkson et al, 2020; Nazari et al., 2017); the thematic content of analyst reports (Huang et al., 2018, Dyer et al., 2017); and the relationship between quarterly earnings conference calls and information asymmetry (Bushee et al., 2018). On the other hand, the critical search for suitable tools for analysing business documents has also expanded. Studies have shown that there is still plenty of room for further research to develop more sophisticated methods and improve the accuracy of results. An example is the work of El-Haj et al. (2019) in which they eaminate four tools for improving contextual

meaning analysis in the business environment. Specifically, they offer significant potential for accounting and finance researchers adopt a forward-looking perspective on the tools and applicability of text mining.

The increasing scientific productivity of text mining in accounting and finance requires a continuous distillation of research on the topic. Therefore, this study conducts a bibliometric analysis with the aim of assessing the state of the art of text mining research in accounting and finance from the two seminal contributions by Loughran and McDonald (2011; 2016). I use sources (i.e., journals) to focus the review. This research aims to examine diverse knowledge published in journals, with different levels of impact, to identify key issues that promote the use of text mining methods in accounting and finance and other related disciplines. In fact, the research questions are as follows:

RQ1) How has the use of text mining techniques in economics and finance evolved?

RQ2) What are the main themes that promote the use of text mining methodologies in accounting and finance?

RQ3) What are the main themes that promote the use of text mining methodologies in other disciplines related to accounting and finance?

Therefore, the main theoretical contribution of this paper is to analyze the development of themes for different levels of source impact and for different disciplines. This provides an opportunity to predict the impact of hot topics on accounting and finance research and to compare it with that of other disciplines.

2.3 Methods

In the analysis of existing publications, bibliometric analysis is an effective strategy. Bibliometrics is the set of methods that can be used to quantitatively analyse academic literature and its changes over time. In fact, it has two main purposes (Noyons et al., 1999), research performance evaluation and scientific mapping. The first purpose is the most common. Performance analysis aims to assess the productivity and impact of different scientific entities (papers, authors, sources). In contrast, scientific mapping aims to show the conceptual, social or intellectual structures of knowledge, as well as its evolution and dynamic aspects (Gutierrez-Salcedo et al., 2018). Bibliometrics is a decisive tool for the future directions of a research field (Chen al. 2017).

Bibliometric studies are particularly valuable in fields such as life sciences and biomedicine. However, bibliometric publications are also emerging in areas that do not have traditional links to this methodology, such as the arts and humanities.

2.3.1 Data Collection

The Web of Science database was systematically searched to identify all publications related to text mining in accounting and financial research. In particular, I queried 2 databases, such as SCI-EXPANDED and SSCI. These are the most widely used databases in the social sciences (Zupic and Čater, 2015) and, unlike other databases, have standardised fields that require little data cleaning.

As an exploratory study, I use the following keywords: 'text*mining' and 'text*analysis'. I only extracted English-language papers on this topic, published in international journals. I focused on the period 2011-2022, as I believe that Loughran and McDonald's work from 2011 is a seminal paper. This query returned 20,692 publications.

I applied some exclusion criteria. Firstly, I only selected publications from the areas of business, corporate finance and management (19,287 records are excluded). Secondly, there were various types of documents, but I decided to consider only articles and procedural documents (125 documents are excluded). This type of document is considered "certified knowledge" with peer review (Ramos-Rodríguez and Ruiz-Navarro 2004).

Figure 3 illustrates the systematic procedure I adopts to arrive at my final dataset of 1,280 publications.

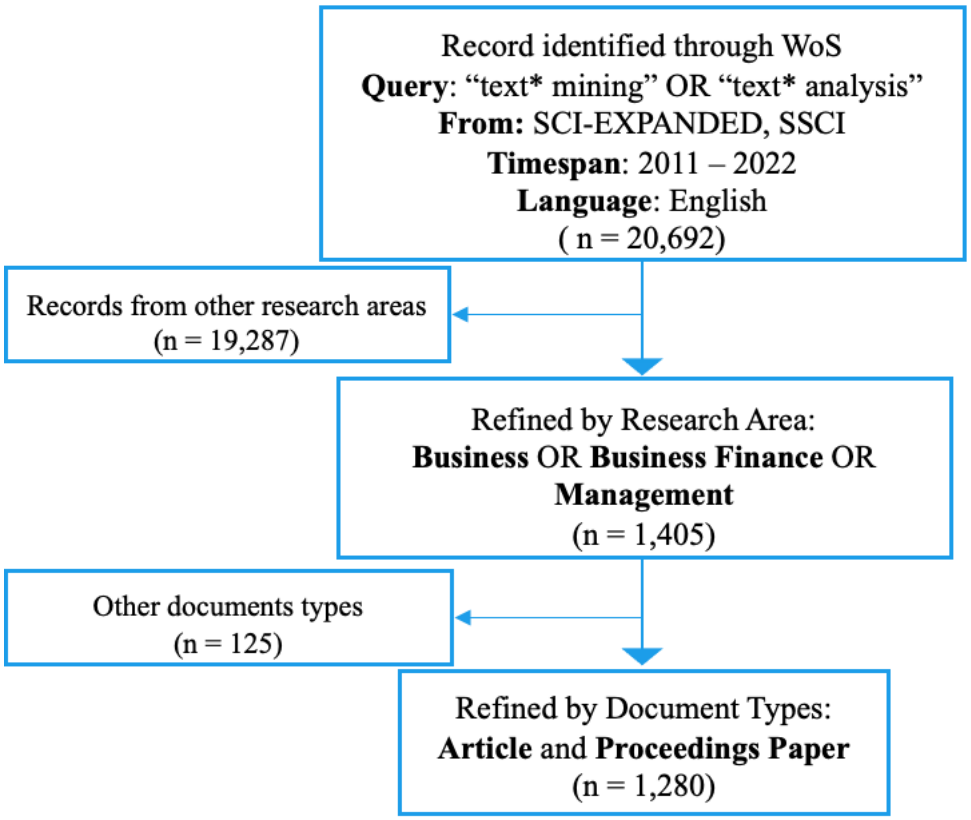


Figure 3 Data collection workflow

2.3.2 Data analysis

In my study, I used the Bibliometrix R package (4.1.1 version) as a bibliometric software tool (Aria, Cuccurullo, 2017). This tool allows us to follow the entire bibliometric analysis workflow from importing bibliographic data from various databases, including Clarivate Analytics' Web of Science, to performing bibliometric analysis and data visualisation. It is an open-source tool for quantitative research in scientometrics and bibliometrics, and today is considered the most comprehensive, integrated and easy-to-use bibliometric tool (Belfiore et al., 2022; Moral-Muñoz et al., 2020).

We then extracted the conceptual structure of text mining for each subsection. This bibliometric analysis helps to generate clusters that allow a broader view of divergent research in a specific scientific field (Börner et al., 2000). The methodological foundation of the conceptual framework is the idea that the co-occurrence of key terms (author keywords, journal keywords, indexing keywords such as ISI WoS Plus keywords or any combination thereof) describes the content of documents in a dataset

(Callon et al., 1991). The more keywords two documents have in common, the more similar the two publications are and the more likely they are to come from the same research field or research speciality at a higher level (Cobo et al., 2011). Keywords, which appear together in a paper, will be related in a word network. By applying Walktrap's clustering algorithm on the co-word network, it is possible to highlight the different themes of a given domain. This thematic network is plotted on a two-dimensional matrix, where the axes are a function of the Callon centrality and Callon density of the thematic map. Centrality can be read as the relevance of the theme in the entire field of research. Density is understood as a measure of the degree of development of the theme. The two-dimensional matrix defines four quadrants concerning different themes (Callon et al., 1991; Coulter et al., 1998; Cahlik, 2000; Aria et al, 2020): (i) motor themes (top right quadrant) that are highly relevant and very well developed; (ii) basic themes (bottom right quadrant) that are highly relevant but not well developed; (iii) emerging or declining themes (bottom left quadrant) that are not very relevant and developed; (iv) niche themes (top left quadrant) that are not very relevant but very well developed. Each theme represented in the two-dimensional matrix is a network cluster. The bubble name is the term, belonging to the cluster, with the highest occurrence value. The size of the bubble is proportional to the occurrences of the words in the cluster. Using the Bibliometrix tool, the thematic evolution analysis used 500 Keywords Plus (ID), with the minimum cluster frequency per thousand documents set to 10 and the inclusion-weighted index for word occurrences set to 0,1. For better clarity and interpretability of the results, I showed 1 keyword per cluster. After an initial analysis, I reformulated some keywords to ensure greater clarity of the final map.

Finally, in line with the two seminal works by Loughran and McDonald, I decided to further divide my collection into two periods (2011; 2016). By dividing the time interval into slices, it is possible to study and trace the evolution of the topic (in terms of its trajectory over time). It is also possible to highlight the tendencies of certain topics to merge, or of a topic to split into several topics, over time.

2.4 Findings

2.4.1 Descriptive results

Figure 4 shows the annual output of scientific publications concerning text mining in accounting and finance since the two seminal contributions by Loughran and McDonald (2011; 2016).

In the first period (2011-2016), the number of articles increased slowly, with a maximum of 61 articles in 2016 alone. This illustrates the early stageof the research. Then, in the second period (2017-2022), the number of articles increased further, peaking in 2022 with 321 articles. This can be called the development stage of the research. A growth rate of 28.7% was recorded in the entire period, but the average age of papers in the collection is 4 years, which indicates that academic production is growing very rapidly especially in recent years.

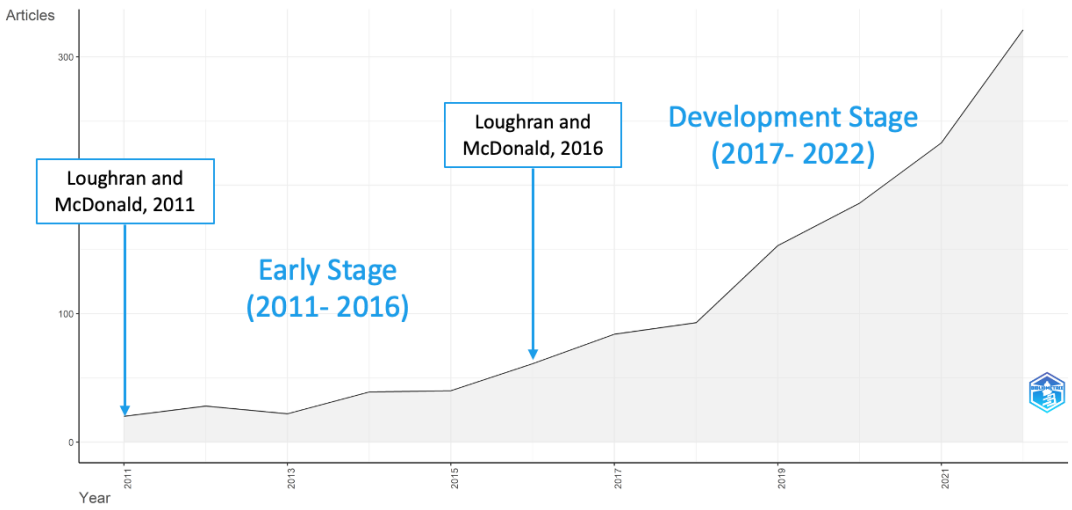


Figure 4 Annual Scientific Production

Papers I collected refer to different economic and financial categories. As Figure 5 shows, the most populated categories are accounting, with 211 articles, and finance, with 196 articles. In addition, the collected articles are published in 291 different sources. Therefore, I divide my collection into 4 sub datasets based on subject categories, distinguishing between publications in accounting and finance or other

related categories, and sources, distinguishing between published in high- or low-impact journals according to the ABS2021 journal ranking.

As can be seen from Table 1, my collection is now divided into 4 subsections: (A) 31 high impact sources from accounting and finance, published 167 articles, (B) 45 low impact sources from accounting and finance, published 240 articles, (C) 78 high impact sources from management categories, published 511 articles, and (D) 137 low impact sources from management categories, published 362 articles.

Table 1 Classification of categories according to the ABS2021 journal ranking

Source impact	**Accounting and Finance categories**	**Management categories**
high (4*, 4, 3)	(A) 31	(C) 78
low (2, 1, other)	(B) 45	(D) 137

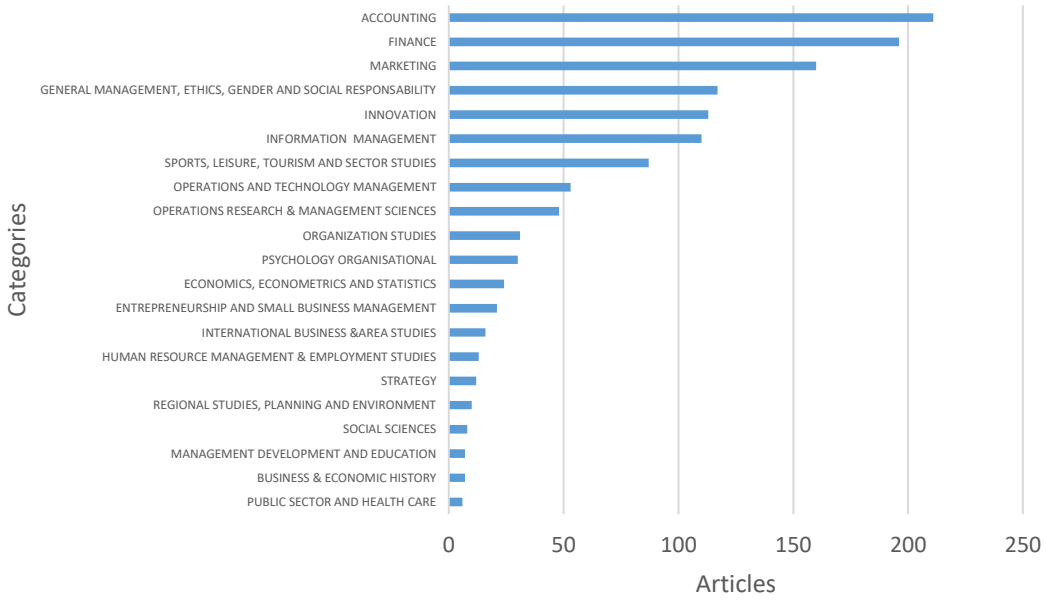


Figure 5 Subject categories

2.4.2 Conceptual structure: thematic evolution

Below are the results of the analysis of the thematic evolution of publications for the four subsections considered. Looking at Callon's centrality and density, I analyse the motor themes, core themes, emerging or declining themes and niche themes. Comparing the thematic map over the two periods considered (2011-2016, 2017-2022), I analyse and trace the evolution of the themes (in terms of trajectory over time).

2.4.2.1 (A) A&F/High Impact

In the first sub-period (Figure 6, upper), I identify six main themes in the category of high-impact accounting and finance publications (A). Among the motor themes I have ‘investors sentiment’ and 'disclosure'. The term ‘investors sentiment’ refers to the analysis of companies' performance and its impact on different sections of the market and investors (Ahmad et al., 2016; Bushman et al., 2016; Ferris et al., 2013). The term 'disclosure' refers to the transparency of information provided by companies to investors regarding the actions taken to attract and manage new investors and the possible legal risks they may face (Elshandidy et al., 2015; Campbell et al., 2014; Loughran and McDonald, 2013).

In the bottom right quadrant, ‘news’ appears as a general theme and encompasses how information is perceived and interpreted by shareholders (Lee, 2016; Sun et al., 2016; Chen et al., 2014). In particular, 'information noise', or an excess of irrelevant information, can make it difficult for shareholders to distinguish between important and unimportant information and this can affect their decisions. For example, the choice of words in a news story can affect the way news is perceived by the public and can influence people's opinions and decisions. On the left quadrant, 'words' and 'firm performance' appear as declining themes as the term ‘talk’ has started to be used more in the field of business studies (Huang et al., 2014; Fiordelisi and Ricci, 2014; Carretta et al., 2011). The latter is confirmed by the presence of the theme ‘talk’ in the upper left quadrant as a niche theme of the period. This theme refers to the analysis of official company statements and comments by analysts and investors concerning company's performance in terms of profits, growth and other key indicators.

In the second sub-period (Figure 6, lower), I identify more topics. Among the motor themes I find topics that refer to the disclosure of information that is not strictly related

to a company's financial performance, such as ‘esg’ and ‘environmental disclosures', but also ‘statements’, which refers to monetary policy announcements, interest rate declarations and other important financial market information by the central bank, and ‘forecast accuracy’, which refers to the communications in which companies discuss their financial results with analysts and investors (Girardone et al., 2021; Bassyouny et al., 2020; Milian et al., 2017). In general, all these words refer to communication and information management by companies.

In this period, ‘monetary policy’, ‘equity’, ‘uncertainty’ appears as general themes. Studies focused on ‘predictability’, ‘distress’, ‘r&d’, ‘wealth’, ‘text’, ‘corporate culture’, which emerged as marginal themes.

In the last quadrant, 'Risk-management' appeared as a very specialised theme of the period.

2.4.2.2 (B) A&F/Low Impact

In the first sub-period (Figure 7, upper), there are no core or niche topics, while there are five cross-cutting topics located between the first and fourth quadrants in the category of low-impact accounting and finance publications (B). One of these includes ‘earnings’ and refers to the communication and management of information about a company's revenues and profits and financial performance, as well as their perception by stakeholders and the media (Frankel et al., 2016; Tetlock, 2014). The other theme concerns understanding and analysing the dynamics of the financial ‘market’, to identify any anomalies or abnormal behaviour that may affect market performance (Price et al., 2012). In the centre of the graph there is a cluster concerning ‘performance’ and it refers to the efficiency with which companies must use their data and knowledge, to identify possible risks and to improve sharing (Sun et al., 2016; Carretta et al., 2015; Andon and Free, 2012).

In the fourth quadrant are two declining themes, one related to ‘science’ and the other to ‘quality’.



Figure 6 (A) Thematic map (time slice 2011-1016; 2017-2022)

In the second sub-period (Figure 7, lower), I identify a plurality of topics. Among the 9 motor themes I find topics referring to ‘decision-making’, ‘investor sentiment’, ‘business risk’, ‘earnings’, ‘big 4’, ‘information-technology’, ‘value relevant accounting information’, ‘disclosures’, and ‘ceo overconfidence’. All these terms refer to the management and transmission of corporate information (Griffith et al., 2020; Wei et al., 2019; Zhou, 2018; Renault, 2017). Information technology is used to support this process, while disclosures and accounting relate to the transparency and communication of this information to external stakeholders. CEO overconfidence and decision-making can affect how information is used within the company, while business risk can be influenced by the quality and reliability of available information. In this period, the evaluation and management of corporate performance appear as general topics. In fact, I find topics such as ‘financial ratios’, ‘sustainability’' and ‘communication’.

Amongst the emerging or declining topics, I find 6 topics concerning 'science', 'discourse', 'directors', 'opportunities', 'innovation' and 'participation'. 'Science' suggest the use of scientific methods and data analysis to better understand corporate activities (Broadstock and Cheng, 2019; Bozanic et al., 2018). 'Discourse' and 'design' suggest the importance of the language and the way of the representation of the activities or the communication of it (Wang and Chen, 2020; Buchholz et al., 2018). 'Directors' suggest the management of business activities through leadership, maintenance, and policymaking. 'Opportunities' and 'innovation' suggest the search for new opportunities and the need for innovation to improve business activities. 'Participation' suggest the importance of understanding and managing business activities within a broader system.

Finally, the influence of political and social decisions such as the use of monetary policy to influence interest rates and money supply appear in the last quadrant as very specialised topics, in fact in the fourth quadrant I find ‘csr’, ‘litigations’ and ‘monetary policy’.

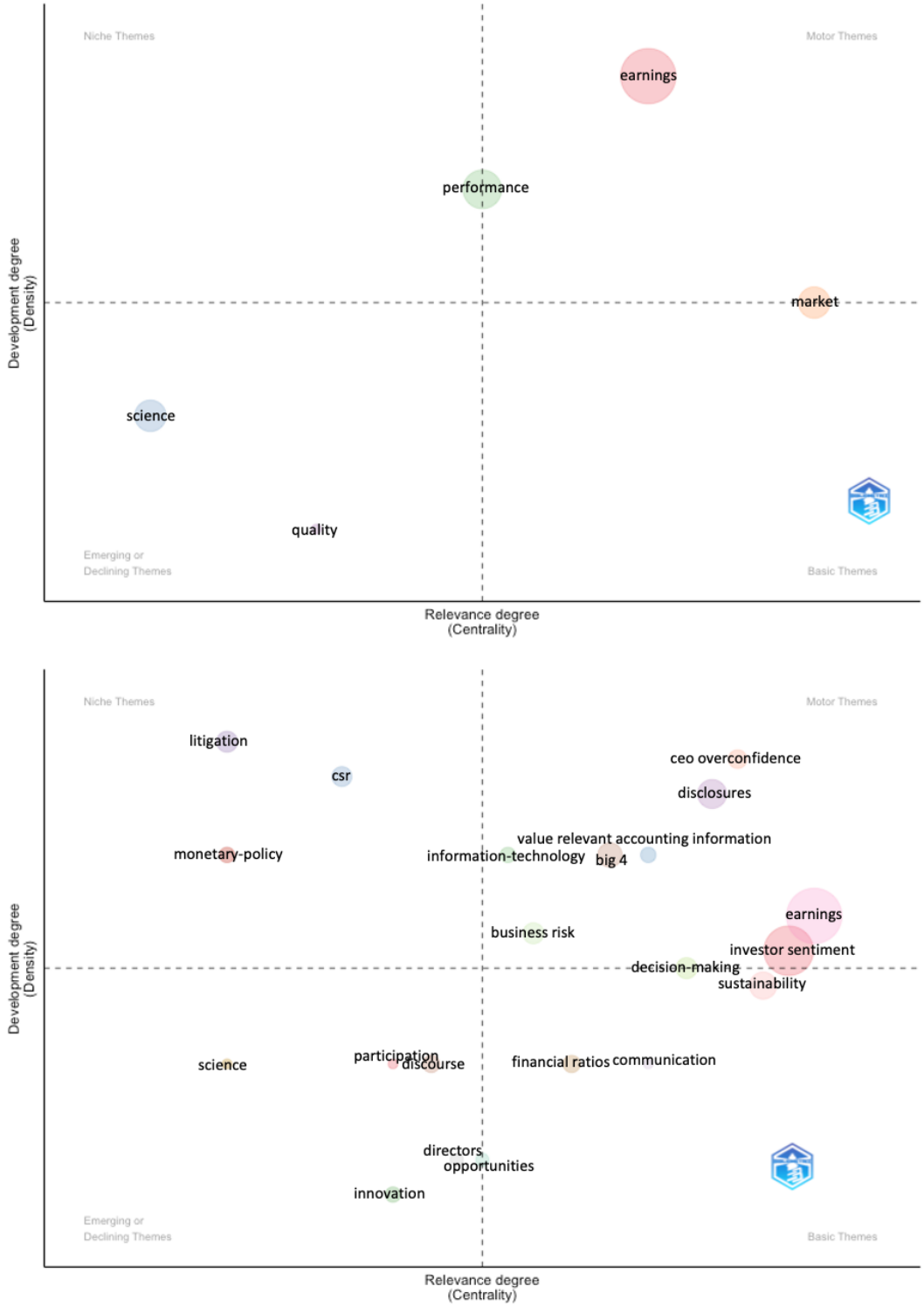


Figure 7 (B) Thematic map (time slice 2011-1016; 2017-2022)

2.4.2.3 (C) Management/High Impact

In the first sub-period (Figure 8, upper), there are a plurality of topics that relate to the management category of high-impact business publications (C). Five clusters are positioned in the first quadrant as motor topics. They include the topics 'consumers', ‘innovation’, ‘discourse’, ‘review’ and ‘framework’ as the most recurring terms. In fact, in these years, is a growing focus on consumer behavior and the impact of culture and national identity on consumption patterns. This has led to an increase in marketing research and the use of textual data to understand consumer demographics and preferences (Crilly et al., 2016; Ludwig et al., 2013). In addition, many papers emphasize the use of bibliometric analysis. In the 'bibliometrics' cluster, I find terms such as artificial intelligence, network structure, and analysis tools. This theme remains a driving force in the second period as the use of bibliometrics is gradually spreading across disciplines. I find terms such as ‘citation’, ‘indicators’, ‘intellectual structure’, ‘intelligence’, and ‘morphological analysis’ and ‘visualization’ referring to the use of textual data to assess the impact and relevance of scientific and academic research (Goh et al., 2013; Ghose et al., 2012). Among the driving themes, research shows a particular focus on the analysis of discourse and communication in the context of organizations. Terms such as ‘perspective’, ‘focus’, ‘rhetorical strategies’, and ‘interpersonal deception’ refer to how language is used in this context to persuade audiences and may relate to the potential for manipulation or distortion of budgets. This theme continues to be a driver in the second period, where the most frequent word becomes ‘disclosure’. Finally, among the driving themes I find ‘framework’, which refers to the need to find a set of common rules, standards and criteria for the classification and interpretation of data extracted from textual documents in order to ensure the consistency and accuracy of the analyses performed. In addition, this framework could also serve as a guide for developing textual analysis methodologies that comply with compliance requirements and accounting standards. The need to find a framework for textual analysis in accounting and finance aims to ensure the quality and reliability of the analyses performed, thereby improving decision-making and forecasting at the firm level.

In the lower right quadrant, ‘aided text analysis’ and ‘innovation’ appear as a basic themes. Effectively, the theme of using artificial intelligence and text analysis

technologies to analyse data on social networks is a general theme. Instead, terms such as design, patents, and productivity suggest an emphasis on the practical application of research and the efficiency of innovation.

On the left side, I find the themes that are still irrelevant and underdeveloped for research such as ‘business’, ‘quality’, ‘indicators’, ‘web’. Among these themes one emerging theme is ‘business’, as this theme becomes a motor theme in the second period (Figure 8, lower). In this theme, I find topics such as ‘performance’, ‘strategy’, ‘decision making’ and ‘competitive advantage’, ‘leadership’, ‘corporate social responsibility’, and ‘entrepreneurial orientation’.

2.4.2.4 (D) Management/Low Impact

In the first subperiod (Figure 9, upper), I identify nine main themes in the publications of the low-impact categories (D). Among the motor themes I have ‘management’, ‘behavior’, and ‘performance’. The term ‘behavior’ refers to people's perceptions and opinions around an enterprise (Kowalczyk et al., 2016). The term ‘performance’ refers to the process of gathering, planning, and using information to support business decisions (Biesenthal and Wilden, 2014; Ghadge et al., 2012; Rossetti et al., 2011).

In the lower right quadrant, ‘innovation’ and ‘organizations’ appear as general themes and relate to the search for new opportunities and the need for innovation to improve the activities of business organizations.

On the left, ‘science’, ‘risk’ appear as declining themes, as these terms do not appear in the later period.



Figure 8 (C) Thematic map (time slice 2011-1016; 2017-2022)

In the second subperiod (Figure 9, lower), I identify a plurality of topics. Among the motor themes I have ‘earnings’, ‘word-of-mouth’, ‘impact’, ‘management’, and ‘networks’. The cluster in which the most frequent word is ‘earnings’ concerns clarity in corporate communication, both in terms of financial reporting and internal and external communication. In fact, I find words such as ‘readability’, ‘tone’, ‘organizational performance’ and ‘corporate governance’. The cluster labelled with ‘word-of-mouth’ contains within it words such as ‘tourism’, ‘hotels’, ‘service quality’, ‘loyalty’, ‘sentiment analysis’, ‘big data’, and ‘analytics’. These words refer to the collection and use of customer data in the tourism industry (Amado et al., 2018; Liao et al., 2018; Birlo et al., 2019). Among the motor themes, I have a cluster that concerns ‘impact’. Looking at the words that make up this cluster I understand that this theme refers to various aspects such as the impact of ‘social media’, ‘text analysis’, ‘value co-creation’, ‘computerized textual analysis’ and ‘meta-analysis’, therefore in general of the impact of the research. The cluster labelled ‘management’ refers to the text analysis aimed at managing the company's image. In this cluster I have terms such as ‘dominant logic’, ‘market opportunities’, and other words concerning the effective management of business activities, performance, and competitive advantages in the market (Ghobakhloo, 2018; Lim and Maglio, 2018). Among the motor themes, I have ‘network’ words in this cluster suggest the use of specific text mining techniques to collect and analyse large amounts of financial and accounting data to support informed decisions.

In the lower right quadrant, ‘orientation’, ‘organizations’, ‘products’, and ‘dynamic capabilities’ appear as general themes.

In the left quadrant ‘brand’, ‘investor sentiment’, ‘resource-based view’ appear as declining themes, while the terms ‘automated text analysis’, ‘news’, ‘popularity’ have started to be used more.

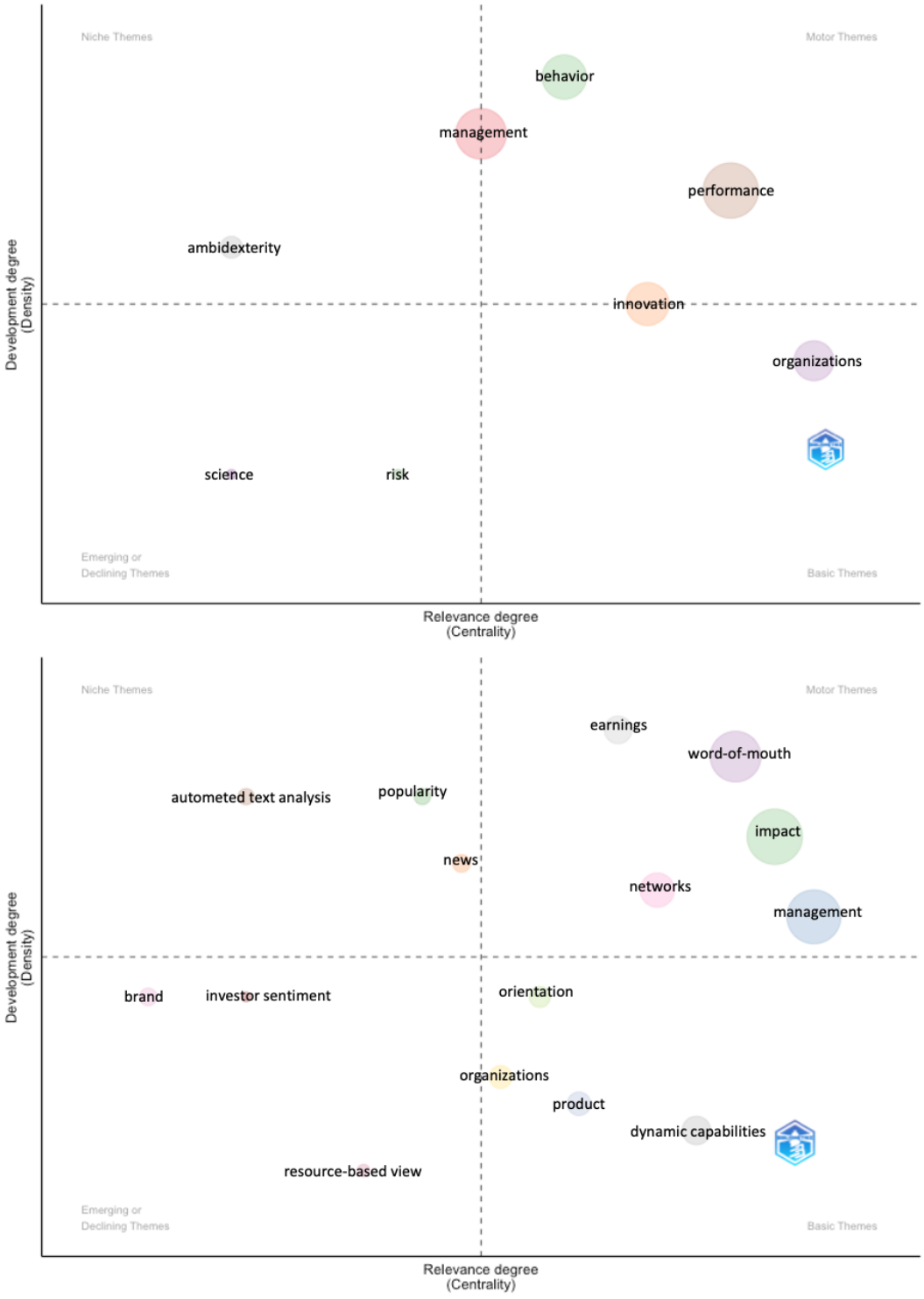


Figure 9 (D) Thematic map (time slice 2011-1016; 2017-2022)

2.5 Conclusions

This study performs a bibliometric analysis of recent and heterogeneous literature on the topic of text mining in economics and finance. The objective is to assess the current state of text mining research in these areas starting with the seminal contributions of Loughran and McDonald (2011, 2016). The study collected articles from 2011 to 2022 and categorized them based on subject categories (accounting and finance or management categories) and source impact (high- or low-impact journals). The study then analyzed the thematic evolution of these publications using Callon's centrality and density, identifying motor themes, core themes, emerging or declining themes, and niche themes.

Several key insights can be drawn from this study. First, the study indicates that text analysis represents a rapidly expanding research field, with an increasing number of publications addressing various current topics and challenges. Second, I note a substantial prevalence within the domains of accounting and finance but the subset of publications in the management category in high-impact journals (C) are the most flourishing. Text mining publications within the accounting and finance categories tend to be published in low impact journals, unlike text mining publications in management, which tend to be published in high impact journals. Third, high-impact accounting and finance publications (A) typically focus on issues related to corporate social responsibility, non-financial reporting, and environmental disclosure, which are becoming increasingly important for businesses as stakeholders demand more transparency and accountability. These publications often address the social and environmental impact of business operations and provide guidance on how to integrate sustainability practices into corporate strategies. On the other hand, low-impact publications in accounting and finance (B) tend to center around more traditional topics such as earnings, accounting information, CEO overconfidence, decision-making, and business risk. While still important areas of study, they may not be as timely or relevant to current business challenges as the high-impact categories. In the management field, high-impact publications (C) often cover bibliometrics, reviews, and online language and communication, as well as textual analysis. These publications focus on understanding how language and communication can shape organizational behavior and outcomes, as well as how to use data and analysis to make

strategic decisions. Finally, low-impact management publications (D), on the other hand, tend to focus on consumer knowledge, word of mouth, and networks. While important for understanding consumer behavior and market trends, these topics may not have as direct an impact on business strategy as the high-impact management categories. Overall, Table 2 shows that high-impact publications in accounting, finance, and management tend to focus on emerging and critical issues that are relevant to current business challenges, while low-impact publications tend to address more traditional and well-established topics.

The contribution of my study, that ranks topics in the accounting and finance field versus the management field, and according to their level of impact and stage of development, can be outlined as follows. First, I used bibliometric methods that make the results reproducible and verifiable. Second, the results of my analysis, which combines publications from different disciplines and with different impact, can be useful to scholars engaged in this field of research and, at the same time, to practitioners in the field of business and management, helping them to stay informed about the latest developments and trends in their area of interest.

Although recent literature has explored the possibilities and limitations of text mining applications in accounting and finance, there is still a lack of interconnection with other disciplines (statistics, linguistics, computer science) that does not allow its full potential to be exploited in accounting and finance. The collaboration between experts from different fields would allow to develop tools and methodologies that can integrate text mining with other data sources such as financial statements, market data, and economic indicators can provide a more comprehensive understanding of accounting and financial data. In summary, there is a need for cross-disciplinary collaboration between experts in accounting, finance, statistics, linguistics, and computer science to fully exploit the potential of text mining in these fields. By working together, it can develop new insights and improve understanding of accounting and finance data.

Table 2 Thematic classification based on categories and the impact of sources

Source impact		Accounting and Finance Categories		
		Themes	**Early Stage**	**Development Stage**
	High	*Motor*	Disclosure, Investor Sentiment	Boards, Corporate Governance, Disclosure, Environmental Disclosures, ESG, Forecast Accuracy, Statements
		Basic	News	Equity, Investor Sentiment, Monetary-Policy, Uncertainty
		Niche	Talk	Methods, Pro Forma Earnings, Risk-Management
		Emerging Declining	Firm Performance, Words	Corporate culture, Distress, Predictability, R&D, Text, Wealth
	Low	*Motor*	Earnings, Performance, Market	Big4, Business Risk, CEO Overconfidence, Disclosures, Decision-Making, Earnings, Investor Sentiment, IT, Value relevant accounting information
		Basic		Communication, Financial Ratios, Sustainability
		Niche		Corporate Social Responsibility, Litigation, Monetary-Policy
		Emerging Declining	Science, Quality	Directors, Discourse, Innovation, Opportunities, Participation, Science
Source impact		**Management Categories**		
		Themes	**Early Stage**	**Development Stage**
	High	*Motor*	Bibliometrics, Consumers, Discourse, Framework, Review	Bibliometrics, Disclosure, Entrepreneurial Orientation
		Basic	Aided Text Analysis, Innovation	Financial Performance, Impact
		Niche	Potential Treatments	Bias, Policy
		Emerging Declining	Business, Indicators, Quality, Web	Energy, Decision, Psychology, Self-Disclosure
	Low	*Motor*	Behavior, Management, Performance	Earnings, Impact, Management, Networks, Word-of-mouth
		Basic	Innovation, Organizations	Dynamic Capabilities, Orientation, Organizations, Product
		Niche	Ambidexterity	Automated Text Analysis, News, Popularity
		Emerging Declining	Risk, Science	Brand, Investor Sentiment, Resource-Based View

2.6 References

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

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). Text summarization techniques: a brief survey. arXiv preprint arXiv:1707.02268.

Amado, A., Cortez, P., Rita, P., & Moro, S. (2018). Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. European Research on Management and Business Economics, 24(1), 1-7.

Andon, P., & Free, C. (2012). Auditing and crisis management: The 2010 Melbourne Storm salary cap scandal. Accounting, Organizations and Society, 37(3), 131-154.

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of informetrics, 11(4), 959-975.

Aria, M., Misuraca, M., & Spano, M. (2020). Mapping the evolution of social research and data science on 30 years of Social Indicators Research. Social Indicators Research, 1-29.

Bassyouny, H., Abdelfattah, T., & Tao, L. (2020). Beyond narrative disclosure tone: The upper echelons theory perspective. International Review of Financial Analysis, 70, 101499.

Bauguess, S. W. 'Use of AI and Machine Learning in Market Risk Assessment'. Remarks at the Practicing Law Institute 2018 SEC Speaks Conference, February 2018.

Belfiore, A., Cuccurullo, C., & Aria, M. (2022). IoT in healthcare: A scientometric analysis. Technological Forecasting and Social Change, 184, 122001.

Biesenthal, C., & Wilden, R. (2014). Multi-level project governance: Trends and opportunities. International journal of project management, 32(8), 1291-1308.

Bilro, R. G., Loureiro, S. M. C., & Guerreiro, J. (2019). Exploring online customer engagement with hospitality products and its relationship with involvement, emotional states, experience and brand advocacy. Journal of Hospitality Marketing & Management, 28(2), 147-171.

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

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. Annual review of information science and technology, 37(1), 179-255.

Bozanic, Z., Roulstone, D. T., & Van Buskirk, A. (2018). Management earnings forecasts and other forward-looking statements. Journal of accounting and economics, 65(1), 1-20.

Broadstock, D. C., & Cheng, L. T. (2019). Time-varying relation between black and green bond price benchmarks: Macroeconomic determinants for the first decade. Finance research letters, 29, 17-22.

Brown, N. C., Crowley, R. M., & Elliott, W. B. (2020). What are you saying? Using topic to detect financial misreporting. Journal of Accounting Research, 58(1), 237-291.

Buchholz, F., Jaeschke, R., Lopatta, K., & Maas, K. (2018). The use of optimistic tone by narcissistic CEOs. Accounting, Auditing & Accountability Journal.

Bushee, B. J., Gow, I. D., & Taylor, D. J. (2018). Linguistic complexity in firm disclosures: Obfuscation or information?. Journal of Accounting Research, 56(1), 85-121.

Bushman, R. M., Hendricks, B. E., & Williams, C. D. (2016). Bank competition: Measurement, decision-making, and risk-taking. Journal of Accounting Research, 54(3), 777-826.

Cahlik, T. (2000). Comparison of the maps of science. Scientometrics, 49(3), 373-387.

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. Scientometrics, 22(1), 155-205.

Campbell, J. L., Chen, H., Dhaliwal, D. S., Lu, H. M., & Steele, L. B. (2014). The information content of mandatory risk factor disclosures in corporate filings. Review of Accounting Studies, 19, 396-455.

Carretta, A., Farina, V., Fiordelisi, F., Schwizer, P., & Lopes, F. S. S. (2015). Don't stand so close to me: The role of supervisory style in banking stability. Journal of Banking & Finance, 52, 180-188.

Carretta, A., Farina, V., Martelli, D., Fiordelisi, F., & Schwizer, P. (2011). The impact of corporate governance press news on stock market returns. European financial management, 17(1), 100-119.

Chen, H., De, P., Hu, Y., & Hwang, B. H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. The review of financial studies, 27(5), 1367-1403.

Chen, X., Chen, B., Zhang, C., & Hao, T. (2017, September). Discovering the recent research in natural language processing field based on a statistical approach. In International Symposium on Emerging Technologies for Education (pp. 507-517). Springer, Cham.

Christensen, H. B., Floyd, E., Liu, L. Y., & Maffett, M. (2017). The real effects of mandated information on social responsibility in financial reports: Evidence from mine-safety records. Journal of Accounting and Economics, 64(2-3), 284-304.

Cieslak, A., & Vissing-Jorgensen, A. (2021). The economics of the Fed put. The Review of Financial Studies, 34(9), 4045-4089.

Clarkson, P. M., Ponn, J., Richardson, G. D., Rudzicz, F., Tsang, A., & Wang, J. (2020). A textual analysis of US corporate social responsibility reports. Abacus, 56(1), 3-34.

Clarkson, P., Li, Y., Richardson, G., & Tsang, A. (2019). Causes and consequences of voluntary assurance of CSR reports: International evidence involving Dow Jones

Sustainability Index Inclusion and Firm Valuation. Accounting, Auditing & Accountability Journal.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. Journal of the American Society for information Science and Technology, 62(7), 1382-1402.

Coulter, N., Monarch, I., & Konda, S. (1998). Software engineering as seen through its research literature: A study in co-word analysis. Journal of the American Society for Information Science, 49(13), 1206-1223.

Crilly, D., Hansen, M., & Zollo, M. (2016). The grammar of decoupling: A cognitive-linguistic perspective on firms' sustainability claims and stakeholders' interpretation. Academy of Management Journal, 59(2), 705-729.

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. Discourse Processes, 54(5-6), 340-359.

Dou, W., Wang, X., Chang, R., & Ribarsky, W. (2011, October). Paralleltopics: A probabilistic approach to exploring document collections. In 2011 IEEE conference on visual analytics science and technology (VAST) (pp. 231-240). IEEE.

Du, S., & Yu, K. (2021). Do corporate social responsibility reports convey value relevant information? Evidence from report readability and tone. Journal of business ethics, 172(2), 253-274.

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.

El-Haj, M., Rayson, P., Walker, M., Young, S., & Simaki, V. (2019). In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse. Journal of Business Finance & Accounting, 46(3-4), 265-306.

Elshandidy, T., Fraser, I., & Hussainey, K. (2015). What drives mandatory and voluntary risk reporting variations across Germany, UK and US?. The British Accounting Review, 47(4), 376-394.

Ferris, S. P., Hao, Q., & Liao, M. Y. (2013). The effect of issuer conservatism on IPO pricing and performance. Review of Finance, 17(3), 993-1027.

Fiordelisi, F., & Ricci, O. (2014). Corporate culture and CEO turnover. Journal of Corporate Finance, 28, 66-82.

Frankel, R., Jennings, J., & Lee, J. (2016). Using unstructured and qualitative disclosures to explain accruals. Journal of Accounting and Economics, 62(2-3), 209-227.

Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: a survey. Artificial Intelligence Review, 47(1), 1-66.

Ghadge, A., Dani, S., & Kalawsky, R. (2012). Supply chain risk management: present and future scope. The international journal of logistics management.

Ghobakhloo, M. (2018). The future of manufacturing industry: a strategic roadmap toward Industry 4.0. Journal of manufacturing technology management, 29(6), 910-936.

Ghose, A., Ipeirotis, P. G., & Li, B. (2012). Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. Marketing Science, 31(3), 493-520.

Girardone, C., Kokas, S., & Wood, G. (2021). Diversity and women in finance: Challenges and future perspectives. Journal of Corporate Finance, 71, 101906.

Goh, K. Y., Heng, C. S., & Lin, Z. (2013). Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content. Information systems research, 24(1), 88-107.

Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. international journal of Computer Applications, 68(13), 13-18.

Griffith, J., Najand, M., & Shen, J. (2020). Emotions in the stock market. Journal of Behavioral Finance, 21(1), 42-56.

Guay, W., Samuels, D., & Taylor, D. (2016). Guiding through the fog: Financial statements complexity and voluntary disclosure. Journal of Accounting and Economics, 62(2-3), 234-269.

Gutiérrez-Salcedo M, Martínez MÁ, Moral-Munoz JA, et al. (2017) Some bibliometric procedures for analyzing and evaluating research fields. Applied Intelligence.

Hearst, M. (2003). What is text mining. SIMS, UC

Hoberg, G., & Lewis, C. (2017). Do fraudulent firms produce abnormal disclosure?. Journal of Corporate Finance, 43, 58-85.

Huang, A. (2008, April). Similarity measures for text document clustering. In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand (Vol. 4, pp. 9-56).

Huang, A. H., Lehavy, R., Zang, A. Y., & Zheng, R. (2018). Analyst information discovery and interpretation roles: A topic modeling approach. Management science, 64(6), 2833-2855.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Naval Technical Training Command Millington TN Research Branch.

Kowalczyk, C. M., & Pounders, K. R. (2016). Transforming celebrities through social media: the role of authenticity and emotional attachment. Journal of product & brand management, 25(4), 345-356.

Lee, J. (2016). Can investors detect managers' lack of spontaneity? Adherence to predetermined scripts during earnings conference calls. The Accounting Review, 91(1), 229-250.

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.

Liao, Q. V., Mas-ud Hussain, M., Chandar, P., Davis, M., Khazaeni, Y., Crasso, M. P., ... & Geyer, W. (2018, April). All work and no play?. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-13).

Lim, C., & Maglio, P. P. (2018). Data-driven understanding of smart service systems through text mining. Service Science, 10(2), 154-180.

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

Liu, B. (2020). Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge university press.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining text data (pp. 415-463). Springer, Boston, MA.

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

Loughran, T., & McDonald, B. (2013). IPO first-day returns, offer price revisions, volatility, and form S-1 language. Journal of Financial Economics, 109(2), 307-326.

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

Ludwig, S., De Ruyter, K., Friedman, M., Brüggen, E. C., Wetzels, M., & Pfann, G. (2013). More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. Journal of marketing, 77(1), 87-103.

Manchiraju, H., & Rajgopal, S. (2017). Does corporate social responsibility (CSR) create shareholder value? Evidence from the Indian Companies Act 2013. Journal of Accounting Research, 55(5), 1257-1300.

Manning, C. D. (2008). Introduction to information retrieval. Syngress Publishing,.

Milian, J. A., Smith, A. L., & Alfonso, E. (2017). Does an analyst's access to information vary with the favorableness of their language when speaking to management?. Accounting Horizons, 31(4), 13-31.

Mohr, J. W., & Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. Poetics, 41(6), 545-569.

Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., & Cobo, M. J. (2020). Software tools for conducting bibliometric analysis in science: An up-to-date review.

Muslu, V., Mutlu, S., Radhakrishnan, S., & Tsang, A. (2019). Corporate social responsibility report narratives and analyst forecast accuracy. Journal of Business Ethics, 154(4), 1119-1142.

Nazari, J. A., Hrazdil, K., & Mahmoudian, F. (2017). Assessing social and environmental performance through narrative complexity in CSR reports. Journal of Contemporary Accounting & Economics, 13(2), 166-178.

Noyons ECM, Moed HF and Luwel M (1999) Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. Journal of the American Society for Information Science 50(2): 115–131.

Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. Journal of Banking & Finance, 36(4), 992-1011.

Ramos-Rodríguez, A. R., & Ruíz-Navarro, J. (2004). Changes in the intellectual structure of strategic management research: A bibliometric study of the Strategic Management Journal, 1980–2000. Strategic management journal, 25(10), 981-1004.

Renault, T. (2017). Intraday online investor sentiment and return patterns in the US stock market. Journal of Banking & Finance, 84, 25-40.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert systems with applications, 33(1), 135-146.

Rossetti, C. L., Handfield, R., & Dooley, K. J. (2011). Forces, trends, and decisions in pharmaceutical supply chain management. International Journal of Physical Distribution & Logistics Management.

Senter, R. J., & Smith, E. A. (1967). Automated readability index. Cincinnati Univ OH.

Sun, A., Lachanski, M., & Fabozzi, F. J. (2016). Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. International Review of Financial Analysis, 48, 272-281.

Sun, L., Najand, M., & Shen, J. (2016). Stock return predictability and investor sentiment: A high-frequency perspective. Journal of Banking & Finance, 73, 147-164.

Tetlock, P. C. (2014). Information transmission in finance. Annu. Rev. Financ. Econ., 6(1), 365-384.

Wang, S., & Chen, X. (2020). Recognizing CEO personality and its impact on business performance: Mining linguistic cues from social media. Information & Management, 57(5), 103173.

Wang, Z., Hsieh, T. S., & Sarkis, J. (2018). CSR performance and the readability of CSR reports: too good to be true?. Corporate Social Responsibility and Environmental Management, 25(1), 66-79.

CHAPTER 3

Financial Configurations of Italian private hospitals: an evolutionary analysis*

ALESSANDRA BELFIORE[1] CORRADO CUCCURULLO♠ MASSIMO ARIA•

Abstract

Private hospitals are an important pillar in many hybrid systems. In Italy they are bound to grow, but we have little knowledge of how they are coping with financial pressures. We use a configurational approach to determine the relationship between the macroprofiles of Italian private hospitals and their performance. We built a unique dataset with governance and financial statements data of all Italian private hospitals. We use a combination of partial triadic analysis (PTA) and clustering technique to identify both the main explicative financial dimensions and hospital configurations and trace their evolutionary paths from the beginning of Regional Health Care Turnaround Plans (2008) to 2016. Understanding the evolution of configurations, our study entails also some implications for policy and practice.

Keywords: Private hospital, PTA, health policy, financial profile, economic performance

---

* This chapter was published on Health Policy in 2022. The regular citation is Belfiore, A., Cuccurullo, C., & Aria, M. (2022). Financial configurations of Italian private hospitals: an evolutionary analysis. Health Policy, 126(7), 661-667. As reported in the copyright section on the Elsevier Health Policy web page, it is possible use and share this works for scholarly purposes (with full acknowledgement of the original article) because chapter this is not published commercially. Copyright policies and co-author permissions are attached to this thesis (https://www.elsevier.com/about/our-business/policies/copyright#Author-rights)

[1] PhD candidate in *Entrepreneurship and Innovation* – Università degli Studi della Campania "Luigi Vanvitelli"- e-mail: alessandra.belfiore@unicampania.it

♠ Full Professor of *Management and Economics* – Università degli Studi della Campania "Luigi Vanvitelli" e-mail: corrado.cuccurullo@unicampania.it

• Full Professor in *Statistics for Social Sciences* – Università degli Studi di Napoli Federico II e-mail: aria@unina.it

3.1 Introduction

In Italy, the private hospital sector is very significant and tends to expand [29]. The accredited hospitals, financed by the National Health Service (NHS), are the majority. They cover 30.4% of the beds nationwide, although a profound heterogeneity at the regional level exists. Since 2007, the incidence of accredited private individuals has been increasing. In the overall reduction in hospital supply, public hospitals cut their beds more remarkably. Most accredited hospitals are small family businesses (60% have less than 100 beds; only 11% exceed 200 beds; 9 out of top 10 private hospital groups belong to families, according to Mediobanca Report 2010-2014) that face an industrial concentration process, due to the difficulties of the public sector in financing health system.

Although, in Italy, there is a vast scientific literature about public hospitals [28], academic concern in private hospitals has risen recently [27, 7, 11], especially about financing, size and levels of industry concentration [11]. To date, no study has dealt about economic performance, financial profile, and structural growth. Beyond the conceptual relevance, there is also a practical one.

Since 2007, various political and economic macro-events have marked the sustainability of the I-NHS. First, the launch of Regional Healthcare Turnaround Plans [15, 31], following the introduction of fiscal federalism in 2001 and the financial responsibility of the Regions to cover their health deficit. Second, the 2008 macroeconomic crisis and the sovereign debt crisis in the EU (2010-12) which resulted in the introduction of rationing measures. Therefore, understanding the economic, financial, and growth profiles over a 10-year period - during a deep and long financial crisis - can be useful to find management and policy implications. We use a configurational approach combining partial triadic analysis (PTA) and cluster analysis to identify Italian accredited private hospitals' financial typologies and trace their evolution over time. Alternative configurations can lead to different paths, which can change over time. Our research questions are the following:

- RQ 1. What financial aspects best explain the heterogeneity between private hospitals?
- RQ 2. What are the main configurations of private hospitals?
- RQ 3. How have these configurations changed over time? (2008; 2012; 2016)

Our paper is organized as follows. Section 2 is a review of the existing literature on private hospitals. Section 3 describes the research methods we used to collect and analyse the data. Section 4 outlines the findings of the analysis. Section 5 discusses the main key information deriving from the analysis and the implications for research, policymakers, and managers. Section 6 reports in a final reflection on our research.

3.2 Background

3.2.1 Current knowledge about private hospitals

Recently, in Italy too, academic and policy interest in the private hospital sectors has been growing. Scholars' perspective tends to focus on the legal form and location [9, 37, 27], on the size (number of beds or wards) [9, 8], on the business model in terms of mix (service mix; payer mix), and on installed and used capacity [9, 8, 11], setting appropriateness of production [8]. Other sources, classified among gray literature, such as the "Associazione Italiana Ospedali Privati" (National Private Hospitals Association) annual reports, focus on the observation of a constantly changing reality, to offer to operators, decision-makers, and public opinion, analysis, and reflections that affect both the ways in which hospital services are carried out and the behavior and opinions of citizens who request services and evaluate their activities.

However, there is a complete lack of studies on the most essential characteristics and configurations of private hospitals and how these have evolved and evolve over the years as a result of major changes in regulation and funding. No study has ever investigated private hospitals and their evolution. The Italian setting is appropriate because since 1992 it has been in continuous institutional evolution as a result of macroeconomic reforms and shocks. These changes modified the regulatory and financing scenario of private hospitals, implying significant strategic and operational choices, in terms of industry concentration, too.

3.2.2 Towards a practical framework

The identification of hospital-specific features and configurations is a necessary step towards designing policy and management interventions to improve their performance. Obviously, there is no single stand-alone action that is effective in improving hospital performance, but a multidimensional bundle of actions tuned to the

context is rather required. Some reviews have appeared over the years to assess the effects on economic, administrative, and healthcare outcomes of different types of property of hospitals [48, 20]. Anyway, hospital performance is influenced by factors that may be wholly or partially out of the control of the hospital management. So, we have two broad categories of impactful variables. First, system-wide factors, which hospitals do not govern. Second, hospital-wide factors, which can be managed by the hospital. We consider hospitals as open systems that must interact with the healthcare system and other ones for survival, adaptation, and growth [32].

3.2.2.1 System-wide factors in Italy

In the early 1990s, a prospective financing system and a broad administrative decentralization were introduced in the Italian NHS, so today there are mainly 21 local systems on regional basis. Each Region has sufficient autonomy to organize its own system. Each of them decides (i) how many health authorities and hospital trusts to establish (ii) how many private hospitals can operate under public funding, (iii) what are the governance bodies of the system, (iv) the reimbursement mechanisms, and (v) budget constraints [23]. Thus, each Region establishes the degree of competition in its quasi-market [16]. For private hospitals, each Region regulates, for example, the number of beds (size) by medical and surgical specialties (scope), organizational, workforce, and technological standards (degree of cost structure rigidity), fees for services (profitability), and budget constraints (asset utilization). As the potential role for competition in healthcare was often mixed [17], starting from the end of the 1990s and for almost the entire first decade of the 2000s, many Regions have reformed their system by mitigating the quasi-market and encouraging collaboration between public providers. A new season began with the introduction of fiscal federalism in Italy (2001) that resulted, from 2007 to 2010, in the Regional Healthcare Turnaround Plans in 10 different Regions [15]. These regions had to reduce growth in healthcare spending and provided a major contribution to reduce financial deficits. Any strategic policy occurred under the supervision of Ministries of Health and Economic Affairs. Among the various measures adopted by the Regions, there are also the regulation of reimbursement fees and new strict predetermined budget constraints. The plans were supposed to last 3 years, but they still exist today. Following the 2008 macroeconomic

crisis, even the Regions without Turnaround Plans played a more effective role in the governance of their system: they limited the autonomy of all providers, they enforced policies similar to recovery plans'.

In summary, in all Regions, regulation has become more stringent with profound repercussions on private hospitals. So, their location strongly influences their strategic, economic, and financial profile. Furthermore, Regions pay providers with different timescales that induce private hospitals to systematically resort to financing operations that modify their financial profile [1, 4, 3, 2, 13, 26, 30].

3.2.2.2 Hospital-level factors

Evidence about the effect of hospital private nature on performance is mixed. As profit-making, private hospitals should achieve profitability under any conditions [40]. When they face greater financial pressure, they tend to improve their efficiency [19] and adjust their financing schemes [10, 35].

Hospital reimbursement mechanisms are important to understand financial implications [25]. In Italy, private hospitals were found to be less efficient than public ones. One reason is that regional specific regulations set a limit to the number of funded admissions and budget constraints are quite volatile [12, 5].

Moreover, most Italian private hospitals are small family businesses. Generally, size (beds) is frequently and positively associated to operating and total margins, but negatively correlated to cash flow [36]. Family ownership has implications on financial configurations of private hospitals [14]. A 2015 Mediobanca survey reports that 9 out of 10 largest Italian hospital groups, in terms of number of employees and turnover, are family businesses and often use control-enhancing mechanisms to reinforce the control, especially cross-shareholdings [41]. Most Italian private hospitals are stand-alone organizations with family control and capital fragmented among numerous second or third generation family members. Sometimes, ownership is enlarged to physicians under debt-to-equity transactions. Generally, in Italian family businesses, equity prevails on debt as financial source, because families do not want to lose control. Indeed, family businesses tend to put socio-emotional values before business values [42]. Many family members are engaged in business and therefore personnel costs are often higher. Moreover, family business is more risk-adverse and

traditional banking finance is the most common source of debt. All these features affect firm ability to sustain the highly rigid hospital cost structure and growth through mergers and acquisitions.

3.3 Materials and Methods

The objective of this study is to analyse the distinct configurations of companies operating in the private hospital sector. By using the PTA, combined with a clustering technique, we map: (i) the different corporate governance characteristics of the Private hospital; (ii) the most relevant economic-financial indicators.

3.3.1 Data collection

We queried Aida database by Bureau van Dijk to find the governance and financial data of all the Italian accredited private hospitals that duly filed the financial statements in the years 2008-2012-2016. The query was based on the private hospitals and nursing homes. We extracted the following information.

- The Region where the operational headquarters are located;
- The ownership structure, such as (i) legal form, (ii) number of shareholders, (iii) BVD independence indicator;
- The hospital size based on (i) employees and (ii) total assets;
- The operational growth (revenues) and structural growth (employees and total assets);
- The economic performance, through (i) ROA, (ii) Ebitda margin, (iii) net income/sales;
- The financial profile is understood as (i) debt/equity ratio, (ii) financial leverage, and (iii) primary liquidity ratio.

The query counted 1,165 firms. We proceeded excluding (Figure 10):

- firms that presented the following legal forms: Consortium (42), Social Cooperative (27), Sole Proprietorship (14), Foundation (3), Limited Partnership (2), Simple Company (2);
- firms that carry out rehabilitation activities and nursing homes, that is to say, those realities that do not provide ordinary or day hospitalization services, as they have financing mechanisms (non-accredited private hospitals);

- firms for which complete information was not available for the three financial years considered.

Our final dataset is composed of 198 companies observed in 3 different periods (2008 - 2012 - 2016).

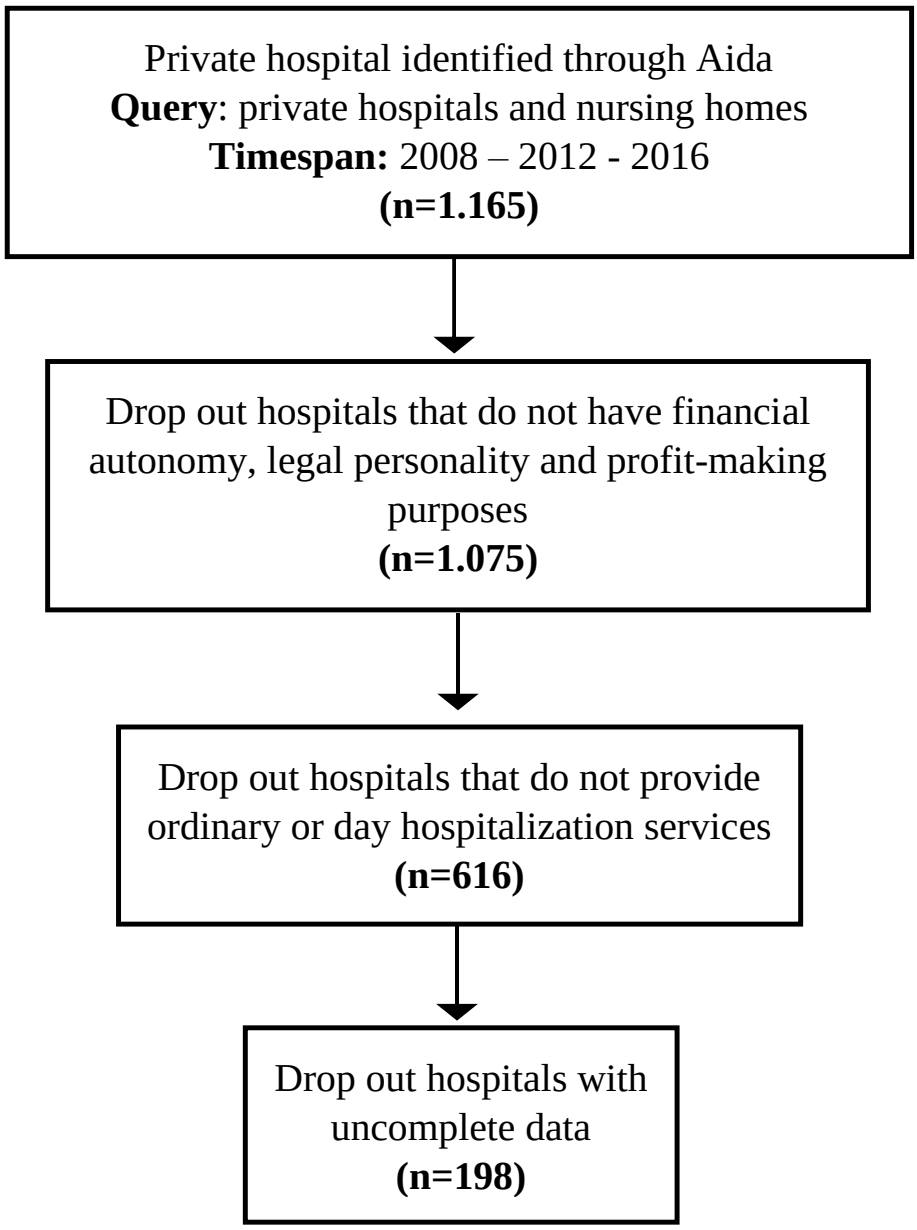


Figure 10 Data collection workflow

3.3.2 Data analysis

3.3.2.1 Partial triadic analysis (PTA)

To answer the first research question and therefore bring out the most differentiating features of the Italian accredited private hospital, we used Partial triadic analysis [21]. PTA, also called X-STATIS, is an extension of principal components analysis (PCA) [22], that is useful for managing multiple data tables that measure sets of variables collected on the same observations but at different times or places [45, 43, 44, 38, 6, 33, 47]. It is a technique based on a simplified approach of three modalities of factor analysis [49, 24] which allows to (i) compare relationships between different datasets, (ii) integrate these datasets into an optimal weighted average called compromise, (iii) project each original dataset on a compromise to analyse commonalities and discrepancies.

Ade4, a multivariate data analysis package for the R statistical environment [46] was used for the analysis. We performed the PTA on our panel data, covering 198 accredited private hospitals, in three different periods. The three tables, which form our historical series, are standardized for each year, and then transformed into a single data frame.

3.3.2.2 Hierarchical clustering

To answer our second research question, that is to identify the configuration of Italian accredited private hospitals, we used the agglomerative hierarchical clustering technique [7]. Since configurational research is based on the assumption that there are distinct and meaningful groupings of companies within a larger entity, grouping methods are central to this part of investigation. Cluster analysis has been used several times both in the hospital sector and other industries. Clustering or group analysis is a set of multivariate data analysis techniques aimed at selecting and grouping homogeneous elements in a dataset. Clustering techniques are based on measures related to the similarity between elements [39]. In the hierarchical analysis of clusters, the segmentation technique is based on the logic of minimizing the distances between the statistical units within the groups and maximizing the distances between the groups. The results are internally homogeneous and externally heterogeneous groups.

The hclust function of the R stats package was used for the hierarchical analysis of the clusters.

3.3.2.3 Configuration evolution

Finally, to answer the third research question, that is, how the configuration of Italian accredited private hospitals changes over time, we combined the results of PTA with the results of the hierarchical analysis of the clusters. We identified the centroids of each cluster and analysed their trajectory over time (2008-2012-2016). We choose to use the centroids because studying the trajectory of all 198 private hospitals would lead to misunderstood results. The centers of gravity represent the "average position" of all the companies that are in a particular cluster, or the arithmetic mean of the coordinates of each company on the axes identified through the PTA. Therefore, they are the best representation of our clusters.

3.4 Findings

3.4.1 Ownership structure of private hospitals

In terms of ownership structure, 85 out of 198 Italian accredited private hospitals are organized as companies; 6 of them have a single shareholder. 113 out 198 hospitals are organized as limited companies; 6 of them with a single shareholder. Up to 50% of private hospitals have between 1 and 3 shareholders. The other 50% have 11 shareholders on average. Moreover, they have concentrated ownership. Indeed, according to the BvD Independence Indicator that categorizes the degree of independence of a company, more than 50% of our sample has 1 shareholder with direct ownership of over 50%.

3.4.2 Discovering main financial dimensions

The first two components summarize 58% of the variance of the original variables. Since we are in a compromise situation, we believe that these two axes (Factor 1 and Factor 2) represent the information contained in the original variables sufficiently. Figure 11 reports the coordinates of the Intrastructure step of the PTA and shows the factorial.

The three-period graphs have the same scale and can be overlaid and compared. The advantage of using the PTA lies in the fact that all points are in the same space, so the two axes have the same meaning in all three graphs. We can interpret the first axis as the financial profile axis and the second axis as the profitability (economic performance). Through PTA, these two aspects have been identified as the best aspects to explain the variability among private hospitals.

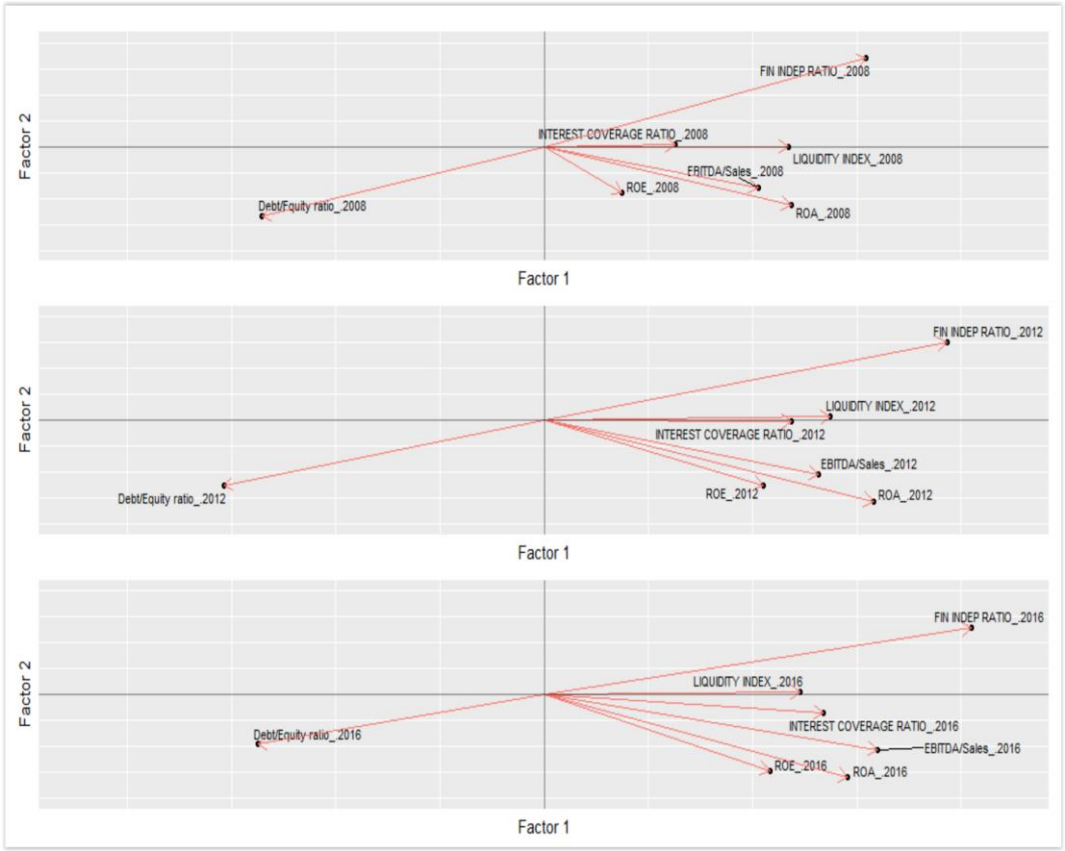


Figure 11 Factorial map of the economic-financial variables

3.4.3 Main financial configurations

Starting from the axes' common structure over the years, it is possible to identify the main configurations of private hospitals, through hierarchical cluster analysis. Figure 12 shows the results of the hierarchical analysis cluster, in the three periods. The 198 private hospitals form 4 different clusters. Looking at the average values of the

variables in each cluster, we can understand what are the configurations of the private hospitals that compose them.

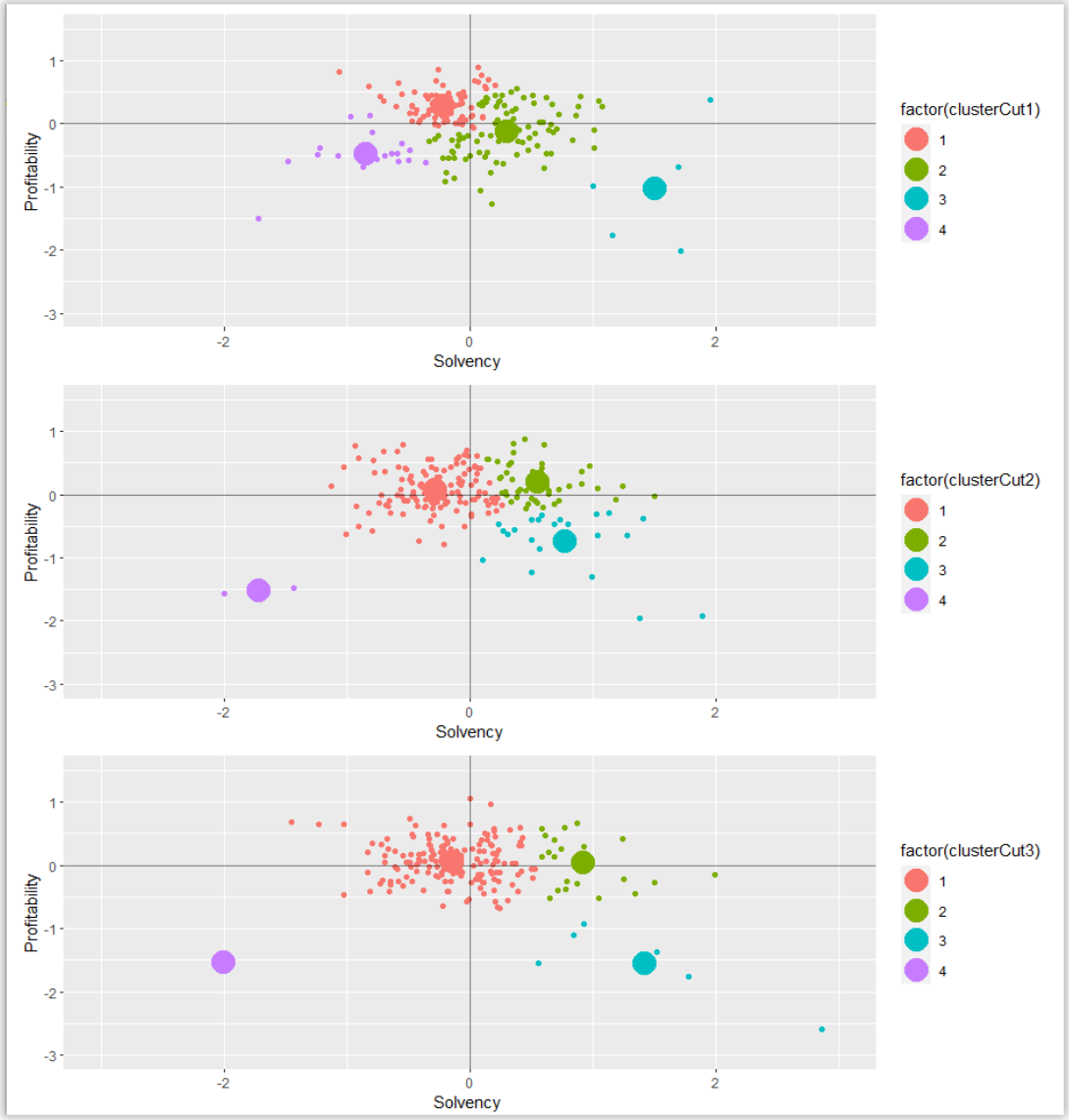


Figure 12 Factorial map of private hospital in 2008, 2012, 2016

1. Cluster 1 is the one located in the upper left quadrant. This cluster reports Italian accredited private hospitals in a situation of severe financial stress. This cluster is defined by a liquidity index and an interest coverage rate of less than 1. The liquidity index is an instrument that expresses the company's ability to meet the financial commitments undertaken. It is given by the ratio between available assets and short-term debt. The numerator represents the amount of

cash in hand and the bank, the readiness for realization, and short-term credits. The denominator is given by the debts to be paid immediately on sight or in the short term. Having a liquidity index lower than 1 means that the companies that are part of this cluster have a shortage of liquidity with respect to short-term debts. The interest coverage rate, on the other hand, indicates the degree of coverage that the operating result can provide at the cost of financial charges. It is given by the ratio between EBITDA and financial charges. The numerator represents the operational management of the company. On the other hand, the denominator is given by financial charges. Similarly, having an interest coverage rate of less than one means that the income generated by the operations is not sufficient to pay off the capital invested to produce it. In this case, the final solvency is negative.

2. Cluster 2 is the one located in the upper right quadrant. This cluster displays Italian accredited private hospitals in an excellent financial and economic situation. In this cluster, we find positive values both of the debt ratios (e.g. liquidity index = 2) and of the profitability ratios (e.g. EBITDA/Sales = 16.5%). In this case, the solvency ability and profitability are positive.
3. Cluster 3 is the one located in the lower right quadrant. This cluster shows Italian accredited private hospitals in a situation of severe economic tension. This cluster is defined by an EBITDA/Sales ratio of less than 10%. The EBITDA/Sales indicator expresses the company's true ability to stay on the market as it measures how much operating income it can generate per unit of turnover. EBITDA is the most important measure of income because it is not influenced by investment policies (through depreciation), financing policies (through interest expense), extraordinary and fiscal policies. Having a low EBITDA/Sales ratio means that businesses are not very profitable. Looking also at the ROE values of private hospitals in cluster 3, we see that the percentage of profitability of invested capital is low. A value tending to zero means that wealth is neither being created nor destroyed. In this case, the final profitability is negative.
4. Finally, cluster 4 is located in the lower-left quadrant. This cluster reports Italian accredited private hospitals in their worst situation. Indeed, in this

cluster we find lower average values, both for debt ratios (eg liquidity index = 0.69) and profitability ratios (ROE = -3.95%). In this case, the solvency and the final profitability are negative.

3.4.4 Tracing configurations' paths

In Figure 13 the coordinates of the centroids for each cluster and period are represented on the PTA plot and the patterns of the clusters over time are highlighted. The size of the spheres reports whether that cluster has grown in number or not, while the arrow shows how it has moved over time.

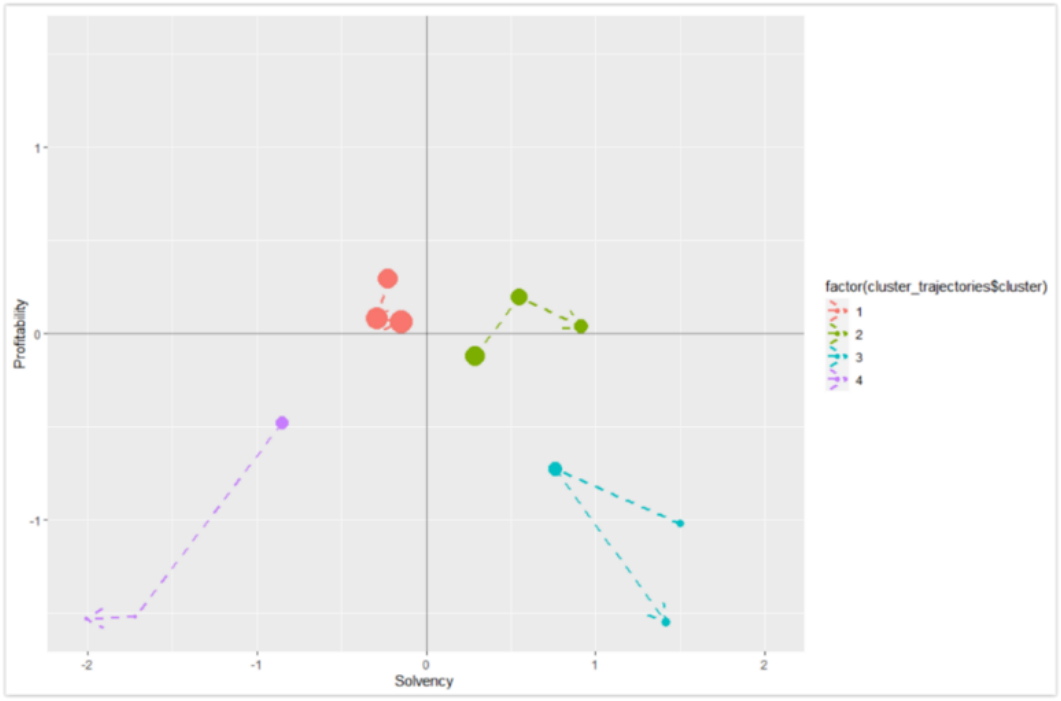


Figure 13 Graphic representation and trajectory of the centroids

First, we can observe the change in size of clusters.

1. Cluster 1, composed of Italian accredited private hospitals with negative solvency, has grown (n = 84 in 2008, n = 130 in 2012, and n = 168 in 2016). Therefore, from 2008 to 2016, the number of private hospitals that became indebted and unable to meet their debts increased.

2. Cluster 2, shrinks from 91 to 44 and then to 22 companies. This cluster represents the best hospitals from both financial and economic points of view. This downsizing highlights the difficulty for these hospitals over the years deriving from the inability of the public sector in financing the healthcare services provided.
3. Cluster 3, made up of those hospitals in serious economic tension, has grown in 2012 but then recovers the initial size in 2016 (n = 5 in 2008, n = 22 in 2012 and n=6 in 2016). This situation shows that the inability of these hospitals to create wealth is only a temporary default as they, as already mentioned, are an important pillar in the Italian healthcare sector.
4. Finally, cluster 4, which represents the worst hospitals from both a financial and economic point of view, downsizes (n = 18 in 2008 n = 2 in 2012-2016).

Instead, as regards the patterns, it is possible to observe that the four clusters move towards different directions over the years.

1. Cluster 1 shifts down from factor 2 in 2012 and then shifts slightly to the right in 2016. This change is due to changes in the values that make up Factor 2, i.e. due to changes in the levels of profitability of firms. Therefore, hospitals in cluster 1, already characterized by financial stress, are then starting to suffer economic stress.
2. Cluster 2 towards Factor 2, in 2012 and then moves straight to the right, towards Factor 1, in 2016. This shift, on both axes, is due to a change in both profitability and solvency values. Therefore, hospitals in cluster 2 continue to improve both financial and economic profiles.
3. Cluster 3 moves upwards in 2012 falling lower than before in 2016. The final result of the displacements of this cluster is a worsening of economic performance. Therefore, hospitals in cluster 3 that already had reduced profitability, see their economic stress increasing.
4. Finally, cluster 4 drops remarkable, over time. This collapse is due to the worsening of both economic and financial profile of the companies that made up this cluster. Therefore, hospitals in cluster 4 continue to deteriorate.

3.5 Discussion

This study provides an overview about the financial configurations of Italian accredited private hospitals and their evolution over time. It provides three relevant contributions. First, the PTA allows to identify solvency and profitability as key financial aspects to explain the heterogeneity among Italian accredited private hospitals. Second, clustering analysis identifies four main financial configurations among these hospitals. Third, we trace clusters evolution paths over time. Our findings show that hospitals in an excellent economic and financial situation have decreased (cluster 2); highly indebted hospitals and weak for profitability have increased (cluster 1); hospitals in worse economic and financial situation (cluster 4) and those less profitable (cluster 3) have worsened their situations. In synthesis, the period coinciding with the macroeconomic crises (2008-16) saw a significant worsening of the financial conditions of Italian private hospitals.

A first research idea for the future is to broaden the timespan of analysis. Probably, Covid-19 will have changed these configurations. A second thread of research is to investigate other characteristics of these hospitals such as structural, operational and governance features. Configurational approach is increasingly applied in strategy, governance, and organization research, but it requires appropriate research methodologies to realize their potential [34]. Although clustering we use in this paper is currently considered the best option to identify configurations, future research could combine it with qualitative comparative analysis to find more interesting and complex configurations [18].

Understanding the evolution of configurations, our study has some implications for policy and practice. Due to the difficulties of the public sector in financing healthcare industry, private hospitals are facing a slow industry concentration. Profitability and debt indicate that Italian private hospitals have non-sustainable business models and are heavily exposed to banks. They need to upgrade to benefit from operational and financial synergies. The industry restructuring process should be encouraged and favored by policy makers with subsidies and incentives. Family businesses increase professionalism in the managerial structure and open equity to non-family members shareholders.

3.6 Conclusion

Italian private hospitals have worsened their financial condition in recent years as macroeconomic crises threatened the sustainability of public healthcare funding. These hospitals need to upgrade and be more complementary to public hospitals.

3.7 References

[1] Amatucci, F. A. B. I. O., & Longo, F. (2009). Quale influenza hanno le partnership pubblico-privato sulla funzione obiettivo delle aziende pubbliche: analisi dei casi in sanità. MECOSAN, 177.

[2] Amatucci, F. M., & Lecci, F. (2006). Le operazioni di partnership finanziaria pubblico-privato in sanità: un'analisi critica, in E. Anessi Pessina and E. Cantù (eds.), L'aziendalizzazione della sanità in Italia. Rapporto OASI 2006, Egea, Milano.

[3] Amatucci, F., Vecchi, V., & Germani, A. (2007). Il project finance in sanità: un'analisi delle caratteristiche economiche e finanziarie. Il project finance in sanità, 1000-1032. in E. Anessi Pessina, E. Cantù (eds), L'aziendalizzazione della sanità in Italia: rapporto OASI 2007, Egea, Milano.

[4] Amatucci, F., Vecchi, V., & Hellowell, M. (2008). Il project finance: quali lezioni dalle esperienze internazionali?. Il project finance, 1000-1028.

[5] Barbetta, G. P., Turati, G., & Zago, A. M. (2007). Behavioral differences between public and private not-for-profit hospitals in the Italian national health service. Health economics, 16(1), 75-96.

[6] Bertrand, F., & Maumy, M. (2010). Using partial triadic analysis for depicting the temporal evolution of spatial structures: assessing phytoplankton structure and succession in a water reservoir. Case Studies In Business, Industry And Government Statistics, 4(1), 23-43.

[7] Bridges Jr, C. C. (1966). Hierarchical cluster analysis. Psychological reports, 18(3), 851-854.

[8] Carbone, C., & Petracca, F. (2015). Gli erogatori privati accreditati: struttura e attività. CERGAS (a cura di), Rapporto OASI, 185-209.

[9] Carbone, C., Longo, F., & Ricci, A. (2014). Il posizionamento strategico degli ospedali privati accreditati: analisi delle leve competitive.

[10] Colombi, R., Martini, G., & Vittadini, G. (2017). Determinanti dell'efficienza ospedaliera transitoria e persistente: il caso dell'Italia. Economia sanitaria , 26 , 5-22.

[11] Cuccurullo, C., & Pennarola, G. (2017). Strategia e performance nelle aziende ospedaliere: analisi, pianificazione e valutazione. EGEA spa.

[12] Daidone, S., & D'Amico, F. (2009). Technical efficiency, specialization and ownership form: evidences from a pooling of Italian hospitals. Journal of productivity Analysis, 32(3), 203-216.

[13] Del Vecchio, M. (2006). Nuove tendenze nelle strutture di governo: interpretazioni e riflessioni a margine di un documento della Regione Emilia Romagna. Nuove tendenze nelle strutture di governo, 1000-1013, in E. Anessi, E. Cantù, Rapporto OASI 2006, Egea, Milano.

[14] Federo, R., Ponomareva, Y., Aguilera, R. V., Saz-Carranza, A., & Losada, C. (2020). Bringing owners back on board: A review of the role of ownership type in board governance. Corporate Governance: An International Review, 28(6), 348-371

[15] Ferrè, F., Cuccurullo, C., & Lega, F. (2012). The challenge and the future of health care turnaround plans: evidence from the Italian experience. Health Policy, 106(1), 3-9.

[16] France, G., Taroni, F., & Donatini, A. (2005). The Italian health-care system. Health economics, 14(S1), S187-S202.

[17] Goddard, M. (2015). Competition in healthcare: good, bad or ugly?. International journal of health policy and management, 4(9), 567.

[18] Greckhamer, T., Furnari, S., Fiss, P. C., & Aguilera, R. V. (2018). Studying configurations with qualitative comparative analysis: Best practices in strategy and organization research. Strategic Organization, 16(4), 482-495.

[19] Hadley, J., Zuckerman, S., & Iezzoni, L. I. (1996). Financial pressure and competition: changes in hospital efficiency and cost-shifting behavior. Medical care, 205-219.

[20] Herrera, C. A., Rada, G., Kuhn-Barrientos, L., & Barrios, X. (2014). Does ownership matter? An overview of systematic reviews of the performance of private for-profit, private not-for-profit and public healthcare providers. PloS one, 9(12), e93456.

[21] Jaffrenou PA. Sur l'Analyse des Familles Finies de Variables Vectorielles: Bases Algébriques et Applications à la Description Statistique, Thèse de Troisième Cycle. Université de Lyon, 1978.

[22] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

[23] Jommi, C., Cantù, E., & Anessi-Pessina, E. (2001). New funding arrangements in the Italian national health service. The International journal of health planning and management, 16(4), 347-368.

[24] Kiers, H. A. (1991). Hierarchical relations among three-way methods. Psychometrika, 56(3), 449-470.

[25] Kruse, F. M., Stadhouders, N. W., Adang, E. M., Groenewoud, S., & Jeurissen, P. P. (2018). Do private hospitals outperform public hospitals regarding efficiency, accessibility, and quality of care in the European Union? A literature review. The International journal of health planning and management, 33(2), e434-e453.

[26] Lecci, F., & Marsilio, M. (2005). Il leasing nelle aziende sanitarie, in E. Anessi Pessina, E. Cantù (eds.), L'aziendalizzazione della sanità in Italia. Rapporto OASI 2005, Milano, Egea.

[27] Lega, F., Petracca, F., & Ricci, A. (2018). Gli erogatori privati accreditati: inquadramento ed evoluzione dei grandi player ospedalieri.

[28] Levaggi, L., & Levaggi, R. (2020). Is there scope for mixed markets in the provision of hospital care? Social Science & Medicine, 247, 112810.

[29] Marotta A., Ricci A. (2021). Gli erogatori privati accreditati: inquadramento e ruolo potenziale nell'implementazione del Piano Nazionale di Ripresa e Resilienza. Rapporto OASI 2021, Egea: Milano

[30] Marsilio, M., & Vecchi, V. (2004). Il finanziamento degli investimenti in sanità, in E. Anessi Pessina, E. Cantù (eds.), L'aziendalizzazione della sanità in Italia. Rapporto OASI 2004, Egea, Milano.

[31] Mauro, M., Maresso, A., & Guglielmo, A. (2017). Health decentralization at a dead-end: towards new recovery plans for Italian hospitals. Health Policy, 121(6), 582-587.

[32] McKee, M., & Healy, J. (2002). The significance of hospitals: an introduction.

[33] Mendes, S., Gómez, J. F., Pereira, M. J., Azeiteiro, U. M., & Galindo-Villardón, M. P. (2010). The efficiency of the Partial Triadic Analysis method: an ecological application. Biometr Lett, 47, 83-106.

[34] Meyer, AD, Tsui, AS, Hinings, CR (1993) Configurational approaches to organizational analysis. Academy of Management Journal 36: 1175–1195.

[35] Mobley IV, L. R., & Magnussen, J. (1998). An international comparison of hospital efficiency: does institutional environment matter?. Applied Economics, 30(8), 1089-1100.

[36] Nurettin Oner, M. S. (2016). Organizational and environmental factors associated with hospital financial performance: A systematic review. Journal of Health Care Finance, 43(2).

[37] Petracca, F., & Ricci, A. (2017). Gli erogatori privati accreditati: attività ospedaliera, distrettuale e valutazione esplorativa degli esiti. CERGAS Bocconi (a cura di), Rapporto Oasi, 257-289.

[38] Rolland, A., Bertrand, F., Maumy, M., & Jacquet, S. (2009). Assessing phytoplankton structure and spatio-temporal dynamics in a freshwater ecosystem using a powerful multiway statistical analysis. Water Research, 43(13), 3155-3168..

[39] Romesburg, C. (2004). Cluster analysis for researchers. Lulu. com.

[40] Rosko, M., Goddard, J., Al-Amin, M., & Tavakoli, M. (2018). Predictors of hospital profitability: A panel study including the early years of the ACA. Journal of Health Care Finance, 44(3).

[41] Saggese, S., Sarto, F., & Cuccurullo, C. (2016). Evolution of the debate on control enhancing mechanisms: a systematic review and bibliometric analysis. International Journal of Management Reviews, 18(4), 417-439.

[42] Sciascia, S., Mazzola, P., & Kellermanns, F. W. (2014). Family management and profitability in private family-owned firms: Introducing generational stage and the socioemotional wealth perspective. Journal of Family Business Strategy, 5(2), 131-137.

[43] Simier, M., Blanc, L., Pellegrin, F., & Nandris, D. (1999). Approche simultanée de $ K $ couples de tableaux: application à l'étude des relations pathologie végétale-environnement. Revue de statistique appliquée, 47(1), 31-46.

[44] Thioulouse, J., Simier, M., & Chessel, D. (2004). Simultaneous analysis of a sequence of paired ecological tables. Ecology, 85(1), 272-283.

[45] Thioulouse, J., & Chessel, D. (1987). Les analyses multitableaux en écologie factorielle. I. De la typologie d'état à la typologie de fonctionnement par l'analyse triadique. Acta oecologica. Série Oecologia generalis, 8(4), 463-480.

[46] Thioulouse, J., & Dray, S. (2007). Interactive multivariate data analysis in R with the ade4 and ade4TkGUI packages. Journal of Statistical Software, 22(5), 1-14.

[47] Thioulouse, J. (2011). Simultaneous analysis of a sequence of paired ecological tables: A comparison of several methods. The Annals of Applied Statistics, 5(4), 2300-2325.

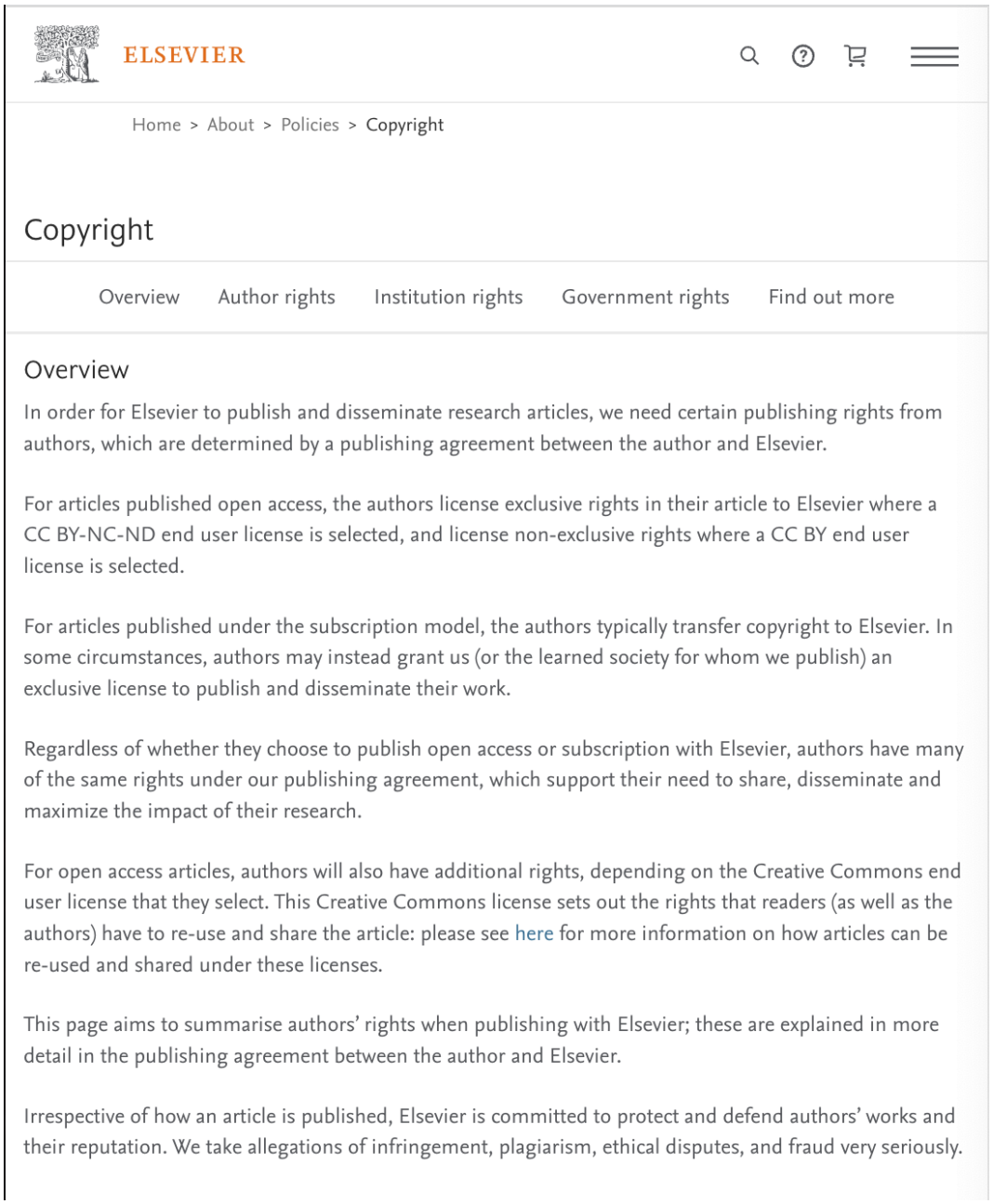
[48] Tiemann, O., & Schreyögg, J. (2012). Changes in hospital efficiency after privatization. Health care management science, 15(4), 310-326.

[49] Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. Psychometrika, 31(3), 279-311.

3.8. Annex

3.8.1 Elsevier Copyright: Free use and share for scholarly purposes, included a thesis

https://www.elsevier.com/about/our-business/policies/copyright#Author-rights



ELSEVIER

Home > About > Policies > Copyright

Copyright

Overview Author rights Institution rights Government rights Find out more

Overview

In order for Elsevier to publish and disseminate research articles, we need certain publishing rights from authors, which are determined by a publishing agreement between the author and Elsevier.

For articles published open access, the authors license exclusive rights in their article to Elsevier where a CC BY-NC-ND end user license is selected, and license non-exclusive rights where a CC BY end user license is selected.

For articles published under the subscription model, the authors typically transfer copyright to Elsevier. In some circumstances, authors may instead grant us (or the learned society for whom we publish) an exclusive license to publish and disseminate their work.

Regardless of whether they choose to publish open access or subscription with Elsevier, authors have many of the same rights under our publishing agreement, which support their need to share, disseminate and maximize the impact of their research.

For open access articles, authors will also have additional rights, depending on the Creative Commons end user license that they select. This Creative Commons license sets out the rights that readers (as well as the authors) have to re-use and share the article: please see here for more information on how articles can be re-used and shared under these licenses.

This page aims to summarise authors' rights when publishing with Elsevier; these are explained in more detail in the publishing agreement between the author and Elsevier.

Irrespective of how an article is published, Elsevier is committed to protect and defend authors' works and their reputation. We take allegations of infringement, plagiarism, ethical disputes, and fraud very seriously.

Author rights

The below table explains the rights that authors have when they publish with Elsevier, for authors who choose to publish either open access or subscription. These apply to the corresponding author and all co-authors.

Author rights in Elsevier's proprietary journals	Published open access	Published subscription
Retain patent and trademark rights	√	√
Retain the rights to use their research data freely without any restriction	√	√
Receive proper attribution and credit for their published work	√	√
Re-use their own material in new works without permission or payment (with full acknowledgement of the original article): 1. Extend an article to book length 2. Include an article in a subsequent compilation of their own work 3. Re-use portions, excerpts, and their own figures or tables in other works.	√	√
Use and share their works for scholarly purposes (with full acknowledgement of the original article): 1. In their own classroom teaching. Electronic and physical distribution of copies is permitted 2. If an author is speaking at a conference, they can present the article and distribute copies to the attendees 3. Distribute the article, including by email, to their students and to research colleagues who they know for their personal use 4. Share and publicize the article via Share Links, which offers 50 days' free access for anyone, without signup or registration 5. Include in a thesis or dissertation (provided this is not published commercially) 6. Share copies of their article privately as part of an invitation-only work group on commercial sites with which the publisher has a hosting agreement	√	√
Publicly share the preprint on any website or repository at any time.	√	√
Publicly share the accepted manuscript on non-commercial sites	√	√ using a CC BY-NC-ND license and usually only after an embargo period (see Sharing Policy for more information)
Publicly share the final published article	√ in line with the author's choice of end user license	×
Retain copyright	√	×

Institution rights

Regardless of how the author chooses to publish with Elsevier, their institution has the right to use articles for classroom teaching and internal training. Articles can be used for these purposes throughout the author's institution, not just by the author:

Institution rights in Elsevier's proprietary journals (providing full acknowledgement of the original article is given)	**All articles**
Copies can be distributed electronically as well as in physical form for classroom teaching and internal training purposes	√
Material can be included in coursework and courseware programs for use within the institution (but not in Massive Open Online Courses)	√
Articles can be included in applications for grant funding	√
Theses and dissertations which contain embedded final published articles as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publication on ScienceDirect	√

Government rights

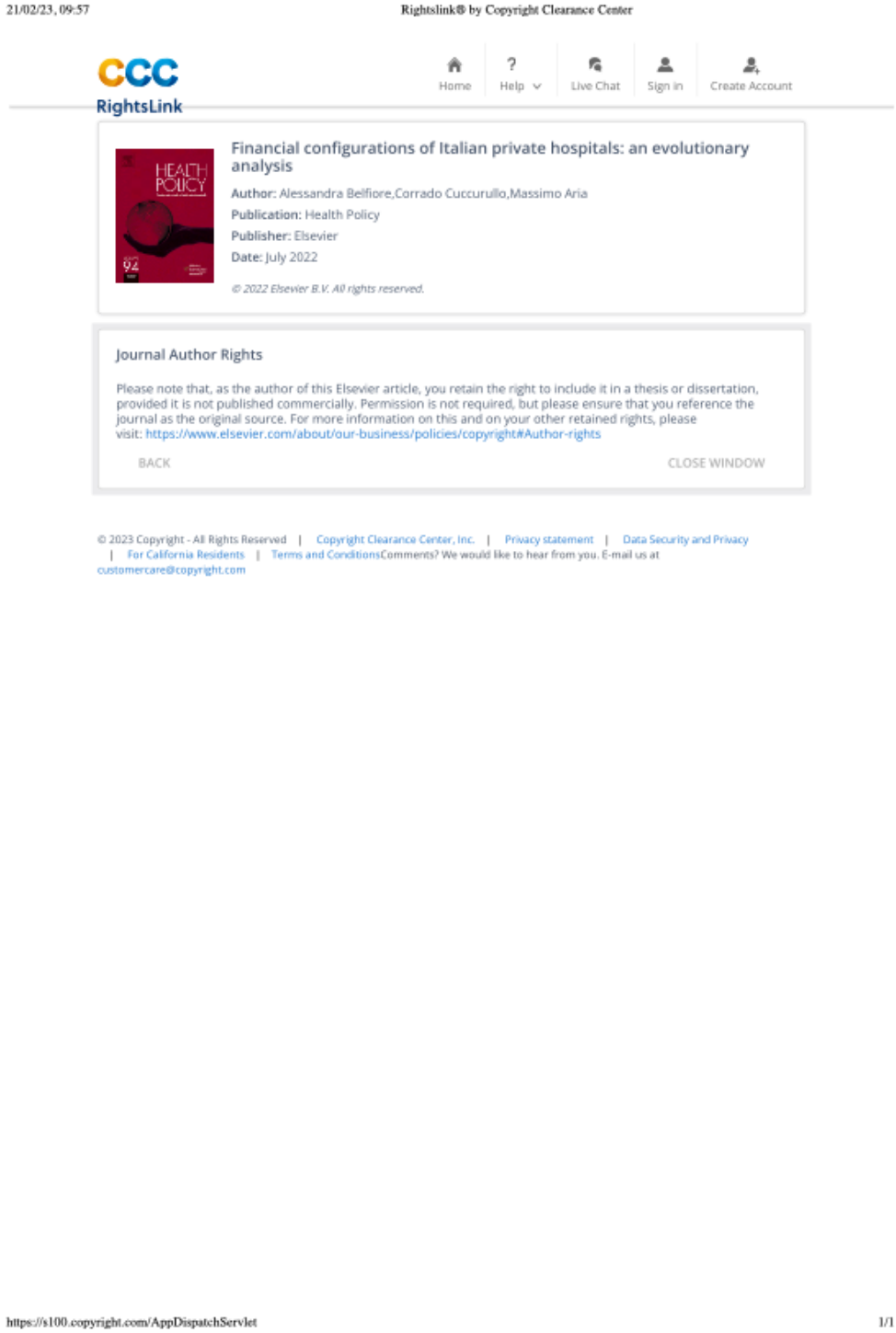
For US government employees, works created within the scope of their employment are considered to be public domain and Elsevier's publishing agreements do not require a transfer or license of rights for such works.

In the UK and certain commonwealth countries, a work created by a government employee is copyrightable, but the government may own the copyright (Crown copyright). Click here for information about UK government employees publishing open access.

Find out more

- Download a sample publishing agreement for articles financed by journal subscriptions in ↓ English and ↓ French.
- Download a sample publishing agreement for articles published open access with a ↓ commercial user license (CC BY) and a ↓ non-commercial user license. (CC BY-NC-ND)
- For authors who wish to self-archive see our sharing guidelines.
- See our author pages for further details about how to promote your article.
- See our hosting page for additional information on hosting research published by Elsevier.
- For use of Elsevier material not defined here please see our permissions page or visit the Permissions Support Center ↗ .
- If an author has become aware of a possible plagiarism, fraud or infringement we recommend contacting their Elsevier publishing contact who can then liaise with our in-house legal department.
- If you are publishing in a society or third party owned journal, they may have different publishing agreements. Please see the journal's Guide for Authors for journal specific copyright information.

3.8.2 RightsLink by Copyright Clearance Center associated to the paper.



3.8.3 Permission of use released by co-authors.

CO-AUTHORS AUTHORIZATION

With this authorization, Corrado Cuccurullo and Massimo Aria, as co-authors, grant the right to use the following scientific article:

Belfiore, A., Cuccurullo, C., & Aria, M. (2022). Financial configurations of Italian private hospitals: an evolutionary analysis. Health Policy, 126(7), 661-667.

in the doctoral thesis.

23/02/2023

Signature



CHAPTER 4

Emotional Language and Financial Performance: Evidence from CEO Discourse during the Covid-19 Pandemic

Abstract

The Covid-19 pandemic has had a significant impact on the way companies communicate with their stakeholders, including customers, employees, investors, and the community at large. Companies have had to adapt quickly to a rapidly changing environment and communicate in new ways to deal with the challenges and uncertainties of the pandemic. This study sought to identify such strategies by performing a textual analysis of letters to shareholders included in the annual reports of companies listed in the Standard & Poors's 500, Euro Stoxx, and Nikkei indexes for the year 2020. A topic modeling procedure and an emotional and sentiment analysis were performed to highlight the key themes and tone of the letters. Then, the correlation between emotions, sentiment, and business performance was examined. This approach, which combines structured and unstructured data, holds promise for the effective analysis of corporate reports because, like financial reports, financial narratives are critically important in determining market trends and influencing investment behavior.

Keywords: topic modeling, emotional analysis, sentiment analysis, communication strategies, annual reports

4.1 Introduction

The unprecedented events of 2020 with the Covid 19 pandemic disrupted not only the normal activities of life but also businesses and economies. The financial results of this difficult year were then disclosed in the spring of 2021, in the company financial statements, adopting appropriate communication strategies. Annual reports are important corporate documents that provide information about the balance sheet and transactions carried out during the previous year. This information is used by shareholders, investors, and analysts to understand the current situation and future direction of the company. In recent years, annual reports have become increasingly important as tools for communicating not only financial performance, but also the strategies and identity of business organizations (Bel et al., 2021; Aerts and Yan, 2017). The online availability of annual reports has enabled the application of advanced technologies to facilitate the analysis and use of this information by analysts. Investors can use this information to make investment decisions, while analysts can use it to assess the company's financial health and future potential (Loughran & McDonald, 2016).

Furthermore, the pandemic has made the transparency and effectiveness of communication by companies in their annual financial statements even more important. Stakeholders wanted to know how companies coped with the crisis and what steps they took to mitigate the negative impacts. In this way, letters to the shareholders, included in the annual reports, have become essential communication tools for companies to present their activities and strategies to their stakeholders, including customers, suppliers and communities (Bondi, 2004; Garzone, 2004; Grove Ditlevsen, 2016; De Groot et al., 2016).

In this context, this contribution aims to identify the communication strategies adopted by companies, in times of general emergency, to convey positive or negative economic results to their stakeholders (Carvache-Franco et al., 2023; Behl et al., 2023). To this end, a single dataset was created, which includes qualitative and quantitative information relating to companies listed on the Standard & Poors' 500, Euro Stoxx and Nikkei indices, relating to the year 2020. Using topic modeling and emotional analysis, the present work aims to detect the hidden thematic structure in the letters to shareholders, identifying the recurring clusters of related words and the use of

emotions and affective language by CEOs to communicate the financial results of their companies. In addition, a Pearson correlation coefficient analysis was conducted to examine whether the affective and emotional values of the letters correlated with the firm's total return.

The work is organized as follows. Section 2 is a review of the existing literature on textual analysis of letters to shareholders. Section 3 describes the research methods used for data collection and analysis, while Section 4 outlines the results of the analysis. Section 5 discusses the main key insights derived from the analysis, as well as the theoretical and practical implications, concluding with a final reflection on the research.

4.2 Literature Review

Textual analysis of letters to shareholders is an important methodology used in many finance and accounting studies to understand the communication practices of companies and evaluate their financial performance. It is an approach that allows you to extract valuable information from companies' annual reports (Glodd and Hristova, 2023; Murphy, 2013). This analysis makes it possible to understand how companies communicate their financial results to their stakeholders, as well as to identify the communication strategies adopted by companies to convey their messages (Belfiore et al., 2022; Pollach, 2012; Prasad and Mir, 2002). Textual analysis combines applications of quantitative and qualitative methods to detect the thematic structure and language used in letters to shareholders, in order to identify recurring themes and emotions expressed by CEOs. Scholars use analysis techniques such as word frequency distribution, topic modeling, emotional and sentiment analysis, and network analysis to examine financial texts such as letters to shareholders.

Word frequency distribution consists of counting the words used in a letter and their ratio to the total number of words in the letter (Poole, 2016; Murphy, 2013). This method makes it possible to identify the most frequent terms and therefore to identify the priorities of the companies, their concerns and the communication strategies adopted.

Topic modeling is a natural language processing technique that allows you to identify the main themes present in financial documents. This method uses advanced

algorithms to analyze the syntactic structure of the text and identify the most representative words and phrases of a given topic (Huang et al., 2018; Törnberg and Törnberg, 2016; Anoop et al., 2016; Pareschi and Mollona, 2016). In this way, it is possible to identify the areas of greatest interest to the company and evaluate their impact on the market.

Emotional and sentiment analysis, on the other hand, focuses on evaluating the emotional attitude expressed in the letter. This method uses natural language processing techniques to identify words and phrases that express positive or negative emotions. In this way, it is possible to understand whether the company is optimistic or pessimistic about the future and to evaluate the effect that this attitude can have on the perception of customers and investors. This technique can be useful in understanding how companies try to influence shareholder perceptions based on the emotion expressed by CEOs in their communications (Ahangari and Sebti, 2023; Costola et al., 2023; Ebrahimian et al., 2023).

Network analysis focuses on evaluating the relationships between the words used in the letter (Callon et al. 1991; Cobo et al., 2011). This method uses network analysis techniques to identify words that are most frequently used together, and which therefore form an important semantic network within the text. In this way, it is possible to understand the relationships between the company's ideas and strategies and evaluate their impact on the market.

In the present study, I use word frequency, topic modelling, emotion analysis and sentiment analysis which are just some of the techniques used in text analysis (Short et al., 2010; Wilkerson and Casas, 2017). Using these techniques, the present work aims to detect the communication strategies of the letters to shareholders, identifying the main topics and the use of emotions and affective language by CEOs to communicate the financial results of their companies. In this way, textual analysis of letters to shareholders become an important source of information for stakeholders, who seek to use this information to make informed, data-driven decisions.

Effectively, the research questions to be addressed are as follows:

RQ1) What communication strategies have companies adopted to communicate their financial results to their stakeholders?

RQ2) How do CEOs use emotion and sentiment language to communicate their companies' financial results?
RQ3) Is there a correlation between the affective and emotional values of letters to shareholders and total return?

4.3 Methods

4.3.1 Data Collection

The sample companies are drawn from the Standard and Poor's 500, (or S&P 500 ®), Eurostock (or Euro Stoxx) and Nikkei indexes, updated to January 2022.
The S&P 500 is a stock market index that tracks the performance of large companies listed on stock exchanges of the United States (Cao et al., 2020). It is widely regarded as the best single indicator of large-cap US stocks. The index comprises 500 leading companies and covers approximately 80% of the available market capitalisation.
The Euro Stoxx index is an equity index representing the performance of the 583 largest listed companies in Europe by market capitalisation. The companies included in the Euro Stoxx index come from 12 European countries, including France, Germany, Italy, Spain and the Netherlands. The Euro Stoxx index is used as a reference for investors wishing to evaluate the performance of the European stock market.
The Nikkei Index is a Japanese stock index representing the performance of the 225 largest companies listed on the Tokyo Stock Exchange based on market capitalization. The Nikkei Index is widely regarded as one of the leading indicators of the Japanese stock market and is used by investors as a benchmark for the performance of listed companies in Japan.
For the preliminary study, I selected all companies listed in the indexes, with reference to the year 2020. The query counted 1308 companies. For each company, the following quantitative and qualitative indicators were extracted. From the annual financial statements, it was possible to recover the text of the letter to the shareholders of the Chief Executive Officer or of the BoD, the headquarters, the sector and the industry in which the company operates. Instead, through the performance of the share, fixed on the closing prices of Bloomberg®4 - at 31 December 2019 and at 31 December 2020 - I define the annual Total Return. It includes dividends, capital gains, payments and

distributions, and any other form of return, and conveys a complete view of investment performance. It measures how much the investment has grown over time and is expressed as a percentage of the original investment.

We proceeded by excluding the companies for which complete information was not available for the years considered. My final dataset consists of 998 companies (352 Standard and Poor's 500 index, 506 Euro Stoxx index and 140 extracted from the Nikkei index).

4.3.2 Data Analysis

4.3.2.1 Topic Modeling

Topic modeling is basically a general term for a collection of methods and algorithms that uncover the hidden thematic structure in document collections by detecting recurring clusters of co-occurring words. There are several different algorithms for performing topic modeling, in this study I use the Latent Dirichlet allocation (LDA) (Blei et al., 2003). Topic modeling is mainly used to automatically and inductively structure the text to find recurrent patterns in the documents thus identifying the broader discursive fields. To perform LDA, texts have to be preprocessed following a set of steps well established in the literature on the statistical analysis of text (Wiedemann, 2016).

We used *udpipe* and *topicmodels*, two libraries available in R. Udpipe library to tokenise texts, transform them into lowercase, and to lemmatize them in order to reduce the high variability of the vocabulary (Wijffels, 2023). Topicmodels library to for the Latent Dirichlet Allocation (LDA) model (Grün and Hornik, 2022).

Following a bag of words scheme, at the end of preprocessing I obtain a documents x terms table T(*nxp*), where the generic element t_{ij} is the frequency of the word j in the document i. Then, I remove from the vocabulary the words that appear less than 5 times. This was done to improve the quality of my model and speed up training.

One of the main limitations is that topic modeling requires setting the number of topics to be detected beforehand. To optimally determine the number of topics, I used the *ldatuning* package, which uses 4 metrics to select the right number of topics for the LDA model (Nikita, 2020). For this, I create many LDA models with different values for the number of topics and choose the one that provides the value that minimizes

Arun et al., 2010 and CaoJuan et al., 2009 and that maximizes Deveaud et al., 2014 and Griffiths and Steyvers, 2004, i.e. a number that will generate meaningful and interpretable topics.

4.3.2.2 Emotional analysis and Sentiment analysis

After having determined the discursive and semantic fields of the proposed concepts, I opt for sentiment and emotional analysis of their discursive nature within the environment of my texts. Sentiment and emotional analysis is an important area of classification research in which supervised and unsupervised methods are often used to classify text in an orderly fashion (from negative to positive, for example) rather than categorically. In this study, I use sentiment and emotion as communicative dimensions to explore how they are implemented within my dataset to investigate how affective vocabularies construct themes and representations, solidify conceptual spaces and emphasize "atmosphere" that the speaker intends to create. For the purposes of this study, I opt for the NRC Affect Intensity Lexicon (Mohammad, 2018) which contains 20,000 annotated English terms for intensity scores (actual values) corresponding to 8 basic emotions (Plutchik 1980; Ekman 1992) and 6 lists of Loughran and McDonald sentiments created by examining word usage in a business context (Loughran and McDonald, 2016). The emotions considered are 8 and refer to disgust, surprise, anger, sadness, fear, joy, expectation and trust while the sentiments are 6 and refer to superfluous, litigious, constraining, uncertainty, negative and positive. Finally, I calculated Pearson's correlation coefficient to examine whether the sentiments and emotional scores of the letters were related to the companies' Total Return—that is, whether being profitable actually raises or lowers the emotional and sentimental level of CEO speeches. Due to the mixed type of my categorical and numerical variables, I used the Gower distance (Gower, 1971). In this way it was also possible to insert the variable relating to the Total Return which is dichotomous. The distance assumes a value between 0 (identical) and 1 (maximally dissimilar).

4.4 Findings

The results are described and interpreted in three subsections. First, I present a descriptive analysis of the dataset. Second, I present a textual analysis of the

shareholder letters. Finally, I measure the degree of correlation between corporate performance and the documents emotion/sentiment.

4.4.1. Descriptive findings

Figure 14 shows the number of companies that are part of the different sectors in the S&P 500, the Euro Stoxx and the Nikkei. Overall, there are a total of 12 industries ranked by Bloomberg, which is the leading source of financial information and analysis tools for financial professionals worldwide. In the S&P 500, the Industrial sector is the most represented, with 83 companies, followed by the Financial sector with 63 companies and the Technology sector with 50 companies. In the Euro Stoxx, the Industrial and Financial sector are once again the most represented, with 93 companies, followed by the Consumer Discretionary sector with 59 companies.

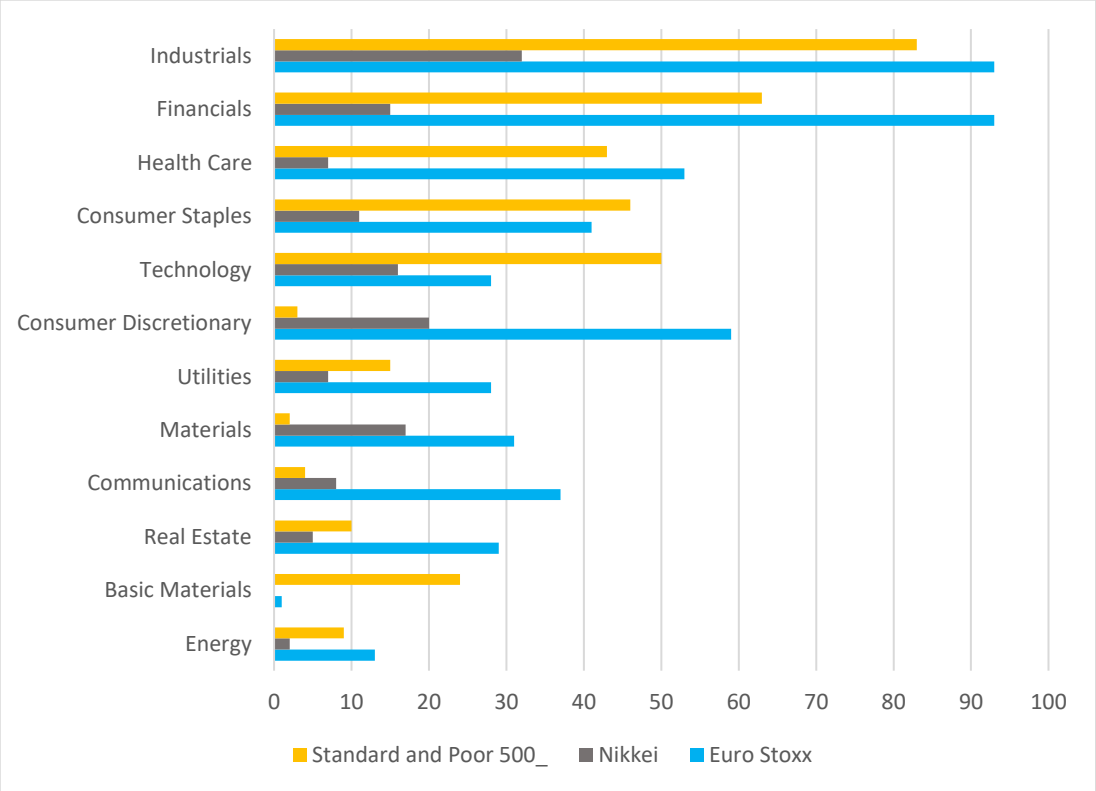


Figure 14 Comparison of sectors in global stock markets

Also in the Nikkei, the Industrial sector is the most represented, with 32 companies, followed by the Consumer Discretionary sector with 20 companies and the Materials

sector with 17 companies. There are significant differences in the representation of different sectors across the three markets. For example, the Technology sector is much more represented in the S&P 500 than the Euro Stoxx and the Nikkei, while the financial sector is much more represented in the Euro Stoxx than the other two markets. These differences reflect the specific characteristics of each area's economy and stock market.

Table 3 shows the number of companies that recorded positive or negative total returns in three major stock indexes - S&P 500, Euro Stoxx and Nikkei - during the period 2019-2020. The Total Return takes into account both the price trend of the asset and the cash flows generated by the asset itself. The number of companies listed as "positive" achieved a positive total return, meaning that these companies achieved price gains and/or dividends from the asset, while the number of companies listed as "negative" had achieved a negative total return, meaning that these companies suffered price losses and/or did not receive any dividends during the reporting period. In the period 2019-2020, the majority of companies in the S&P 500 achieved a negative total return, indicating that it may have been a more volatile index during this period. The number of companies with negative total returns is 218 (62%) for the S&P 500, 241 (48%) for the Euro Stoxx and 47 (34%) for the Nikkei. This suggests that US companies may have been hardest hit during the time of the pandemic.

Table 3 Comparison of financial performance of companies in global stock markets

Total Return (2019-2020)	**S&P's 500**	**Euro Stoxx**	**Nikkei**	
positive	134 (38%)	265 (52%)	93 (66%)	**492**
negative	218 (62%)	241 (48%)	47 (34%)	**506**

Table 4 provides some insight into the lexical and structural features of my dataset.

The texts of the 998 letters to shareholders consist of 64,335 sentences, which on average contain 72 tokens each. The two measures of lexical diversity, the Guiraud Index (29) and the Type-Token Ratio (TTR 0.02) suggest that the dataset contains many repeated words, indicating a narrow focus in the subject matter of the documents.

Also, the average sentence length is relatively short, indicating that the documents are written in a clear and concise style. However, the average document length in characters and tokens is relatively long, suggesting that the dataset contains information-rich documents.

Table 4 Information on the lexical and structural characteristics of the dataset

Description	**Values**
Documents	998
Tokens	1.541.159
Dictionary	35.655
Lemmas	38.994
Sentences	64.335
Docs Avg Length in Chars	289.384
Doc Avg Length in Tokens	1.544
Sent Avg Length in Tokens	72
Sent Avg Length in Chars	186
TTR	0,02
Guiraud Index	29

4.4.2 Topic context

Topic modeling allows us to identify the main topics of the collection by detecting groups of recurring words. The ldatuning package implements 4 metrics to select the perfect number of topics for the LDA model. To identify the number of topics it is enough to find the value that minimizes Arun et al., 2010 and CaoJuan et al., 2009 and that maximizes Deveaud et al., 2014 and Griffiths and Steyvers, 2004. From the graph (Figure 15) it can be concluded that the optimal number of topics is 14. I decided to show the first 10 words with the highest probability of belonging to the topic, for a better readability of the findings (Figure 16).

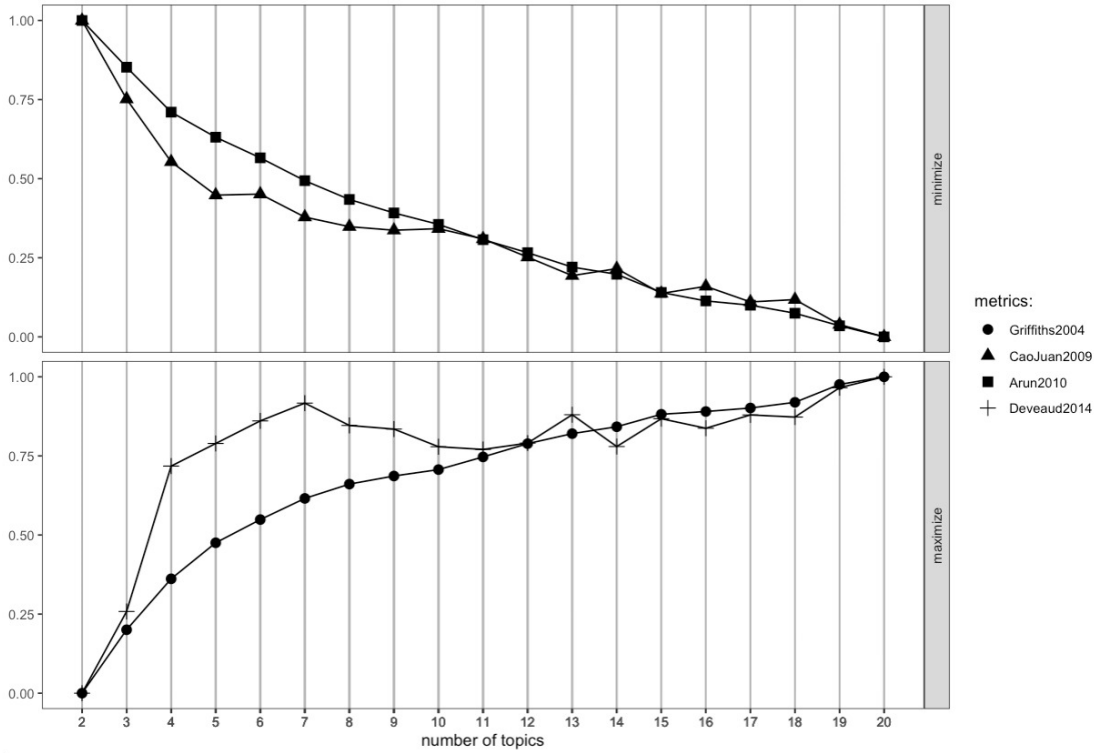


Figure 15 Elbow Plot

Topics 1- 3 - 9 resulting from the analysis of the letters from the CEOs are closely linked to the theme of management and company management. In particular, concepts such as "board", "group", "shareholder", "director" and "supervisory" are mentioned, suggesting a strong focus on corporate governance. Furthermore, the presence of terms such as "year", "crisis", "dividend", "result" and "euro" indicates attention to the company's economic performance and the management of its investments.

Topic 2 is focused on the theme of the global crisis caused by the Covid-19 pandemic. The common trend between the words "COVID", "pandemic", "patient", "healthcare", "vaccine", "life", and "care" is the need to deal effectively with the crisis to protect the world economy. The COVID-19 pandemic has put a strain on economic systems around the world, with negative consequences for the global economy. However, efforts to contain the spread of the virus, treat infected patients and develop effective and safe vaccines have demonstrated the ability of science and innovation to tackle the toughest challenges.

Topics 4 – 6 - 13 highlight the importance of investing in people, processes, and technologies to build a better world. The common trend among the words "people",

“colleague”, “team”, "time", "year" and "lockdown" is the accelerating effects of the fourth industrial revolution. Automation and artificial intelligence have changed the nature of work, while digital technologies have created new opportunities to improve people's quality of life. However, to take full advantage of these opportunities, you need to invest in people, skills and infrastructure. The COVID-19 pandemic has accelerated this transformation, leading to an increase in remote working and the digitalization of business processes. Even CEOs are aware that the pandemic has accelerated the digital transformation of their companies, but they are also aware of the importance of investing in people, providing them with the skills and resources necessary to adapt to change.

In Topic 5 there are the words "FOX", "network", "platform", "news", "brand", "revenue" and "life". Digital transformation has changed the way people access news and media content. Companies must be able to adapt to these transformations to remain competitive. A winning strategy to do this is the diversification of activities and the creation of a strong brand. In this context, "FOX" represents a company that has been able to exploit this strategy effectively. Originally a traditional television network, Fox diversified into an online streaming platform and expanded its social media presence. This diversification has enabled the company to reach a wider audience and grow sustainable long-term revenue.

Topics 7- 10 - 11 – 14 are about how companies have responded to customer needs during the Covid-19 pandemic. The key terms of the topic include "client", "investment", "share", "shareholder", "technology", "solution" and "innovation", which indicate a strong interest of companies in investments and asset portfolio management, as well as attention to customer relations and shareholder interests. In letters to shareholders, the CEOs gave a long-winded description of the technologies and solutions used to improve the customer experience. Data analysis to personalize the offer of services and improve understanding of customer needs is the basis of change. In addition, many companies have developed new products and services due to the need to change the supply chain and sales’ points. This has paved the way for future growth opportunities for the companies.

Topics number 8 - 12 emerging from the analysis of the letters from the CEOs mainly concern the attention of companies towards sustainability and the reduction of the

environmental impact of their activities. In fact, terms such as "emission", "wind energy", "sustainability" and "renewable" indicate a strong concern of companies and a focus on improving production processes to ensure the sustainability of the products and services offered. This indicates a strong focus on establishing long-term strategies for company growth, suggesting that companies are investing in different renewable energy sources to achieve sustainability goals.

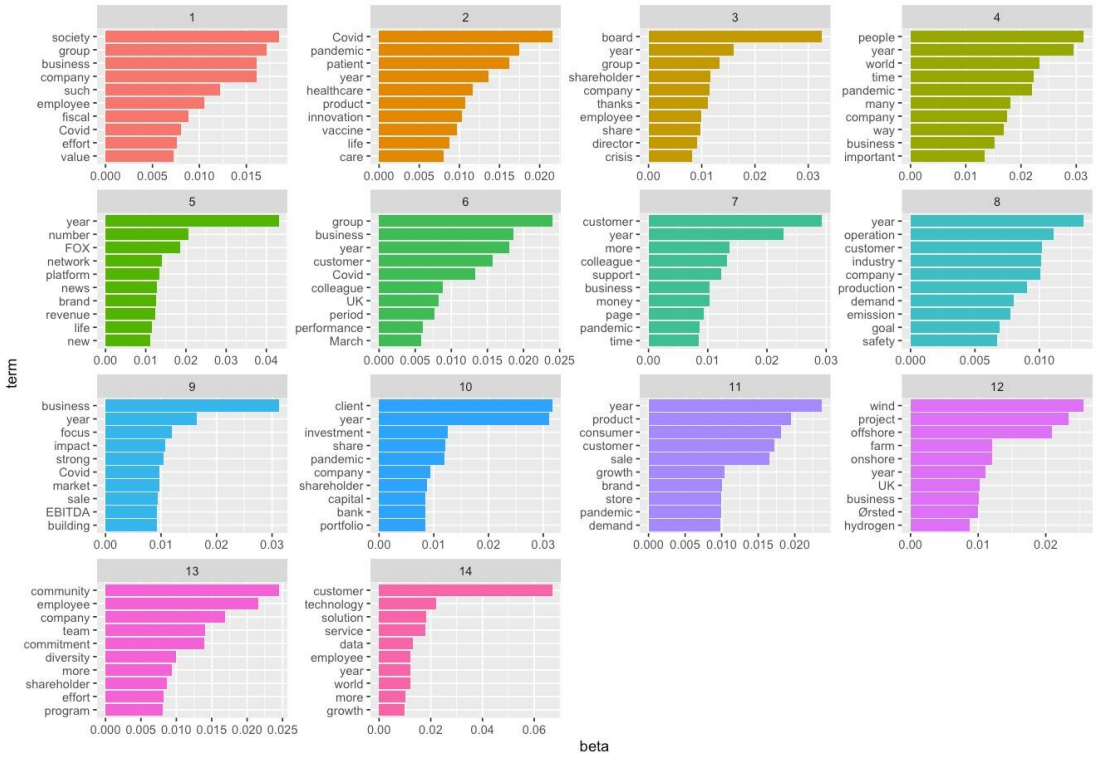


Figure 16. Best Topics

4.4.3 Relationship between emotional/sentimental language and performance

Figure 17 shows the comparison between the results of the analysis of emotions and sentiments on the letters to shareholders.

The prevailing emotions (*Figure 17* upper) expressed in the letters to shareholders are trust (32%), anticipation (25%) and joy (14%). This indicates that CEOs have focused on building trusting relationships with their shareholders, demonstrating a sense of hope and optimism about the future. However, the pandemic has also caused concern and fear about the economic and social consequences, as evidenced by the emotion of fear (11%). Furthermore, the expression of emotions such as sadness (6%), anger (5%)

and disgust (2%) indicate that CEOs also experienced difficulties and frustrations related to the pandemic situation and related business challenges.

The sentiment (*Figure 17* lower) prevalent in letters to shareholders is positive, with 53% of occurrences. However, there is also a high frequency of negative sentiment (25%), suggesting that many CEOs have faced significant challenges during the pandemic. 12% of the words express uncertainty, which reflects the fluid and evolving nature of the pandemic situation. However, 7% of words were classified as "constraining," indicating that CEOs have addressed important limitations due to the pandemic. 3% of the words were classified as "litigious", suggesting that CEOs have not addressed many legal issues during the pandemic. Finally, it is interesting to note that only 1% of the words analyzed were classified as superfluous, which suggests that CEOs have chosen their words carefully and tried to be direct in their communications.

By comparing the results of the analysis of emotions and sentiments, it can be seen that the majority of companies have tried to convey a sense of confidence and stability to investors, trying to mitigate the negative effects of the pandemic on the economy and on the markets. Probably, the CEOs were more concerned about the future and the uncertainty of the economic and political environment, rather than about the stability and solidity of the company.

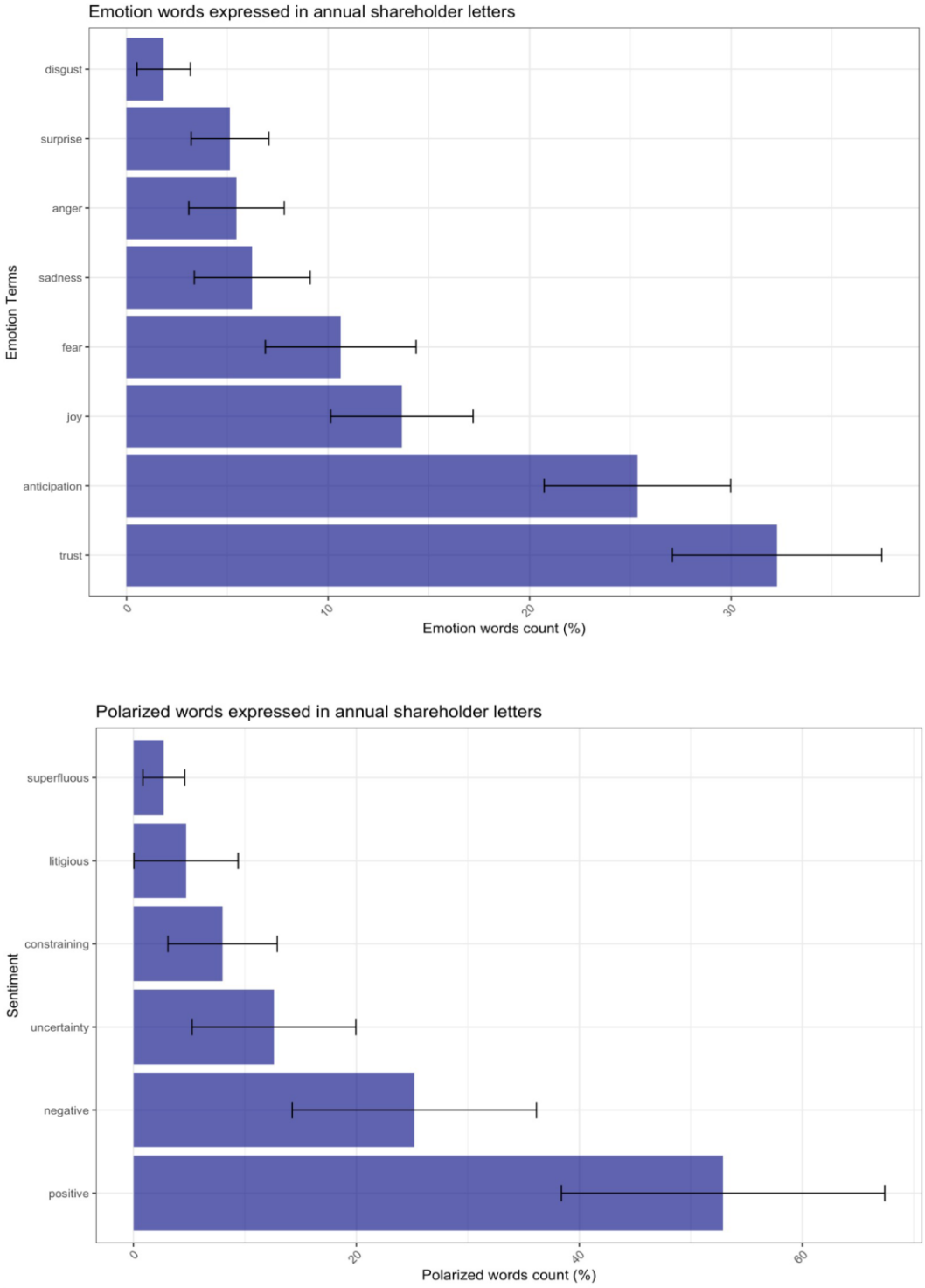


Figure 17 Emotions/Sentiments in letters to shareholders

Pearson's correlation coefficients were used to examine the relationship between the sentiments and emotions expressed in CEO letters and the firm's Total Return. From Figure 18, there are correlations between emotions and sentiments, but there is no relationship between these affective and emotional values with the company's Total Return. There is a very strong positive relationship between Fear intensity score and Sadness intensity score (r = 0.7), Anger intensity score (r = 0.66) and negative feeling (r = 0.63), which suggests that these emotions may be intertwined and exacerbate each other. On the other hand, the intensity of Trust is strongly negatively related to Sadness (r = -0.66) and Anger (r = -0.65), which suggests that trust may play a protective role in reducing the impact of these negative emotions in the CEO's communication. Moreover, the intensity of Anticipation (r = -0.60), Joy (r = -0.57), and Positive sentiment (r = -0.54) is strongly negatively related to fear intensity, which suggests that these emotions may act as buffers or protective factors against fear. Beyond the relationship that exists between the various affective and emotional values in the letters, there is no relationship whatsoever between them and the company's Total Return. It's important to note that correlation does not imply causation. This suggests that the company's financial performance did not significantly influence CEO discourse in terms of emotion or affection. The Covid-19 pandemic was an unexpected shock and, as such, the effects of this extraordinary event are still uncertain and it is not possible to predict the duration, both positive and negative.

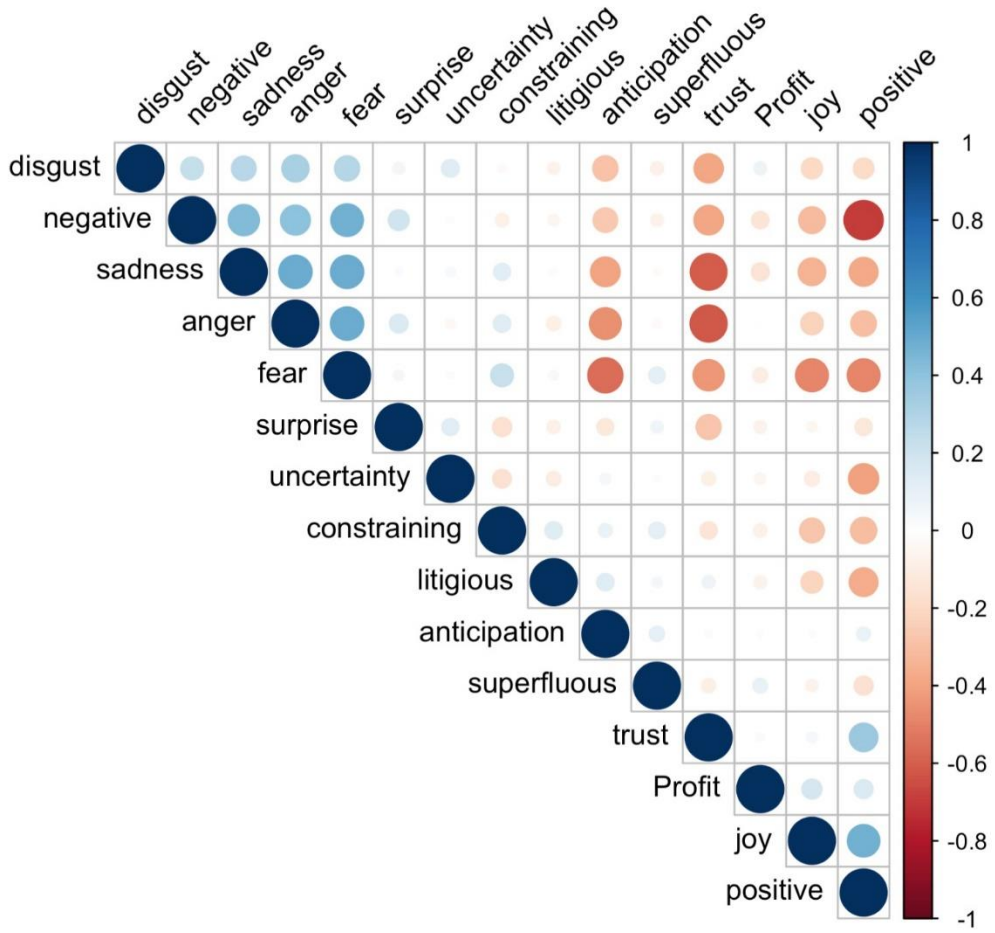


Figure 18 Links among emotions, sentiments, and Total Return

4.5 Conclusions

This study aims to identify the communication strategies adopted by companies during the Covid19 pandemic. The study analyzes the letters to shareholders contained in the annual reports of companies listed in the Standard & Poor's 500, Euro Stoxx and Nikkei indices, for the year 2020. Topic modeling, emotional analysis together with sentiment analysis and analysis of the correlation with Total Return provide a complete view of the strategies and themes adopted by companies in response to the pandemic.

Five key insights can be drawn from this study. First, the companies in my dataset focus on 12 sectors. Among these, the prevalence of the Industrial and Financial sectors stands out. Second, in relation to performance, it was highlighted that most of the companies in the S&P 500 achieved a negative Total Return. This suggests that US companies may have been hardest hit during the time of the pandemic. Third, the textual analysis showed that the letters from the CEOs were rich in information but

written in a clear and concise way. Fourth, the analysis of the letters from the CEOs highlights 14 significant topics. Topics range from corporate governance and economic performance to sustainability, digital transformation, and customer relationships. The COVID-19 pandemic has accelerated many of these transformations, and companies that are able to adapt and invest in people, processes, and technologies will be better equipped to thrive in the post-pandemic era. Fifth, the analysis of emotions and sentiments in shareholder letters during the pandemic period suggests that CEOs have focused on building trust and conveying a sense of hope and optimism for the future, while facing significant challenges. However, looking at the correlation between sentiments and emotions, it can be seen that despite the positive tone, companies have experienced difficulties and frustrations related to the pandemic situation and related business challenges. The CEOs have chosen their words carefully and have tried to be direct in their communications, reflecting a sense of responsibility towards their investors.

A first research idea for the future is to broaden the time frame of the analysis. Probably the effect of the Covid-19 pandemic will have changed these communication strategies also in the years following 2020. A second line of research is to investigate other characteristics of these companies. For example, using structural, operational, and financial indexes it will be possible to identify diversified communication strategies. While the textual approach I use in this paper is currently considered optimal for identifying communication strategies, future research may combine it with other performance information to find more interesting and complex strategies.

By studying the communication strategies adopted by companies during the COVID-19 pandemic, my study has some implications for policy and practice. First, the investigation paves the way for a new paradigm of financial statements analysis that goes beyond the conventional use of ratios and margins to include textual information in financial documents. The combination of massive data with financial forecasting and auditing purposes offers the potential to identify hidden patterns, trends, and relationships within large amounts of financial data. Second, companies are facing rapid sector transformation due to the difficulties caused by the Covid pandemic19. Companies must be able to adapt to these transformations to remain competitive. The process of sector restructuring should be encouraged and facilitated by policymakers

with funding targeted at digitization and sustainability that will enable companies to benefit from greater opportunities for long-term growth and success. At the same time, companies are called upon to review and adapt their investment strategies to meet the challenges posed by the COVID-19 pandemic. This must include greater investment in digital technologies and remote work platforms to ensure business continuity in the event of future crises or disruptions. In addition, companies must consider the importance of maintaining a strong customer orientation and adapting to their changing needs.

4.6 References

Aerts, W., & Yan, B. (2017). Rhetorical impression management in the letter to shareholders and institutional setting: A metadiscourse perspective. Accounting, Auditing & Accountability Journal, 30(2), 404-432.

Ahangari, M., & Sebti, A. (2023). A hybrid approach to sentiment analysis of Iranian stock market user's opinions. International Journal of Engineering, (Articles in Press).

Anoop, Vijayalekshi Sundaresan, S. Asharaf, and P Deepak. 2016. "Unsupervised Concept Hierarchy Learning: A Topic Modeling Guided Approach." Procedia Computer Science 89:386–394. https://doi.org/10.1016/j.procs.2016.06.086.

Behl, A., Jayawardena, N., Nigam, A., Pereira, V., Shankar, A. e Jebarajakirthy, C. (2023). Indagare le strategie di marketing internazionale riviste durante COVID-19 sulla base delle risorse e delle capacità delle aziende: un approccio metodo misto. Giornale di ricerca aziendale , 113662.

Bel, N., Bracons, G., & Anderberg, S. (2021). Finding Evidence of Fraudster Companies in the CEO's Letter to Shareholders with Sentiment Analysis. Information, 12(8), 307.

Belfiore, A., Spano, M., Mandenaki, K., & Giordano, W. Exploring Companies' communication strategies in the Covid-19 era.

Blei D. M., Ng A. Y. and Jordan M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Bondi, M. (2004). "If you think this sounds very complicated, you are correct": Awareness of cultural difference in specialised discourse", in Candlin, Christopher N. / Gotti, Maurizio (ed.) Intercultural Aspects of Specialized Communication. Bern: Peter Lang, 53-78

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. Scientometrics, 22, 155-205.

Cao Juan, Xia Tian, Li Jintao, Zhang Yongdong, and Tang Sheng. 2009. A density-based method for adaptive lda model selection. Neurocomputing — 16th European Symposium on Artificial Neural Networks 2008 72, 7–9: 1775–1781. http://doi.org/10.1016/j.neucom.2008.06.011

Cao, J., Ruan, X., & Zhang, W. (2020). Inferring information from the S&P 500, CBOE VIX, and CBOE SKEW indices. Journal of Futures Markets, 40(6), 945-973.

Carvache-Franco, O., Carvache-Franco, M., Carvache-Franco, W., & Iturralde, K. (2023). Analisi dell'argomento e del sentimento delle comunicazioni di crisi sulla pandemia di COVID-19 negli hashtag turistici di Twitter. Ricerca sul turismo e l'ospitalità , 23 (1), 44-59.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. Journal of informetrics, 5(1), 146-166.

Costola, M., Hinz, O., Nofer, M., & Pelizzon, L. (2023). Machine learning sentiment analysis, Covid-19 news and stock market reactions. Research in International Business and Finance, 101881.

De Groot, E., Nickerson, C., Korzilius, H., & Gerritsen, M. (2016). Picture This: Developing a Model for the Analysis of Visual Metadiscourse. Journal of Business and Technical Communication, 30(2), 165–201.

Ebrahimian, K. A. M. E. L., Abbasi, E., Alam Tabriz, A., & Mohammadzadeh, A. (2023). Stock Price Forecasting with Support Vector Regression Based on Social Network Sentiment Analysis and Technicl Analysis. International Journal of Finance & Managerial Accounting, 8(29), 53-64.

Ekman P. (1992). Are there basic emotions?.

Garzone, G. (2004) Annual Company Reports amd CEO's Letters: discoursal features and cultural markedness, ", in Candlin, Christopher N. / Gotti, Maurizio (ed.) Intercultural Aspects of Specialized Communication. Bern: Peter Lang, 311-341.

Glodd, A., & Hristova, D. (2023). Extraction of Forward-looking Financial Information for Stock Price Prediction from Annual Reports Using NLP Techniques.

Grove Ditlevsen, M. , (2012) "Revealing corporate identities in annual reports", Corporate Communications: An International Journal, Vol. 17 Issue: 3, pp.379-403.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. Biometrics, 857-871.

Grün B, Hornik K (2022). _topicmodels: Topic Models_. R package version 0.2-13, <https://CRAN.R-project.org/package=topicmodels>.

Huang, A. H., Lehavy, R., Zang, A. Y., & Zheng, R. (2018). Analyst information discovery and interpretation roles: A topic modeling approach. Management science, 64(6), 2833-2855.

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

Mohammad S. (2018). Obtaining reliable human ratings of valence, arousal and dominance for 20,000 English words. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 174-184.

Mohammad S. M. (2017). Word affect intensities. arXiv preprint arXiv:1704.08798.

Murphy, A. C. (2013). On “true” portraits of Letters to Shareholders–and the importance of phraseological analysis. International Journal of Corpus Linguistics, 18(1), 57-82.

Nikita M (2020). _ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters_. R package version 1.0.2, <https://CRAN.R-project.org/package=ldatuning>.

Pareschi, Luca and Edoardo Mollona. 2016. “No Country for Neoliberalism: A Topic Modeling Approach to Protean Discourses to Resist Privatizations in Italy.” Institutions and Transition Economics: Firm Governance eJournal. https://doi.org/10.2139/ssrn.2802751

Plutchik R. (1980). A general psychoevolutionary theory of emotion. In Theories of emotion, pp. 3-33. Academic press.

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

Poole, R. (2016). Good times, bad times: A keyword analysis of letters to shareholders of two Fortune 500 banking institutions. International Journal of Business Communication, 53(1), 55-73. https://doi.org/10.1146/annurev-polisci-052615-025542.

Prasad, A., & Mir, R. (2002). Digging deep for meaning: A critical hermeneutic analysis of CEO letters to shareholders in the oil industry. The Journal of Business Communication (1973), 39(1), 92-116.

Rajkumar Arun, V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy. 2010. On finding the natural number of topics with latent dirichlet allocation: Some observations. In Advances in knowledge discovery and data mining, Mohammed J. Zaki, Jeffrey Xu Yu, Balaraman Ravindran and Vikram Pudi (eds.). Springer Berlin Heidelberg, 391–402. http://doi.org/10.1007/978-3-642-13657-3_43

Romain Deveaud, Éric SanJuan, and Patrice Bellot. 2014. Accurate and effective latent concept modeling for ad hoc information retrieval. Document numérique 17, 1: 61–84. http://doi.org/10.3166/dn.17.1.61-84

Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. (2010). Construct validation using computer-aided text analysis (CATA) an illustration using entrepreneurial orientation. Organizational research methods, 13(2), 320-347.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. Proceedings of the National Academy of Sciences 101, suppl 1: 5228–5235. http://doi.org/10.1073/pnas.0307752101

Törnberg, Anton and Petter Törnberg. 2016. “Combining CDA and Topic Modeling: Analyzing Discursive Connections between Islamophobia and Anti-feminism on an Online Forum.” Discourse and Society 27 (4): 401–422. https://doi.org/10.1177/0957926516634546.

Wiedemann G. and Wiedemann. (2016). Text mining for qualitative data analysis in the social sciences (Vol. 1). Wiesbaden, Germany: Springer Fachmedien Wiesbaden.

Wijffels J (2023). _udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit_. R package version 0.8.11, <https://CRAN.R-project.org/package=udpipe>.

Wilkerson, John, D. and Andreu Casas. 2017. “Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges.” Annual Review of Political Science 20 (1): 529–544.

CHAPTER 5

Conclusions

5 Conclusions

5.1 Discussion

Financial analysis is increasingly moving away from the classical view relating only to the analysis of financial statements indicators, using a wide range of advanced tools and technologies to collect, process, and analyze data. The evolution of financial analysis began with the availability of a large amount and variety of data. However, big data and unstructured data require advanced analytical tools to combine and analyze the data effectively. The advent of digital technology combined with the increasing availability of a wide range of data has led to the evolution of financial analysis into a new interdisciplinary field called Informetrics in accounting and finance. This discipline uses data science techniques, computer science, and accounting and finance expertise to collect, process, and analyze large amounts of diverse data. Informetrics in accounting and finance represents a step forward in financial analysis, allowing it to provide a more in-depth and truthful view of companies and the market.

In this thesis, various quantitative methods (bibliometric analysis, conceptual framework of thematic evolution, partial triadic analysis (PTA), hierarchical grouping, theme modeling and emotional analysis) have been applied to different types of data (numerical and textual), extracted from the financial documents of the companies.

In the first study, a bibliometric review was conducted to understand the evolution and issues promoting the use of text mining methodologies in accounting, finance, and management. Three research questions were formulated to guide the analysis, focusing on the evolution and themes of text mining in the various subject categories and for the different levels of impact considered. The Web of Science database was used to identify 1,280 publications which were then broken down by subject categories and source impact. The conceptual structure of the publications was analyzed to reveal the thematic evolution of text mining. The findings suggest that high-impact accounting and finance publications focus on non-financial reporting, while low-impact publications focus on accounting information. Conversely, high-impact publications in management focus on bibliometrics, while low-impact publications focus on consumer knowledge. The implications of the study highlight the opportunities of text

mining methods in different contexts and the possibility of further developments and applications in the considered domain.

In the second study, an evolutionary analysis was conducted to understand the financial configurations of Italian private hospitals. Three research questions were formulated, and the Aida database was used to collect the financial data of 198 accredited Italian private hospitals. PTA, hierarchical clustering and the configurations approach were used to identify four major financial configurations. The results suggest that three of the configurations worsened during the 2008-2016 macroeconomic crisis and that Italian private hospitals have unsustainable business models and are heavily exposed to banks. The implications of the study highlight the importance of defining the essential characteristics of private hospitals and cataloging financial characteristics to provide insightful suggestions to practice and research, especially in the fields of strategy and governance. The study also suggests that configurational studies are a promising development that requires appropriate research methods to realize their potential.

The third study analyzes the communication strategies adopted by CEOs during the Covid-19 pandemic. The study uses data from the Standard and Poor's 500, Euro Stoxx, and Nikkei indices, including CEO text or BOD letters to shareholders and annual total return. The study employs topic modeling and emotional and sentiment analysis to identify communication strategies used by CEOs in their letters to shareholders. The results show that CEOs focused on building trust and in conveying optimism for the future while facing significant challenges, as indicated by the emotion and sentiment analysis and their correlation. The study also identified 14 main topics discussed by CEOs, including corporate governance, economic performance, sustainability, digital transformation, and customer relations. The study finds a correlation between affective and emotional values in letters to shareholders and total return. The study recommends that companies invest in digital technologies and remote work platforms to ensure business continuity in future crisis and maintain a strong customer orientation by adapting to their changing needs. Policymakers should encourage scenario restructuring with funding for digitization and sustainability to help companies adapt and remain competitive. The implications of the study highlight the importance of the words used by CEOs in decision making and provide insights

into the strategic priorities and long-term vision of the company. The study suggests that analyzing emotional language and sentiment in financial documents can provide new insights for investment planning, particularly during times of crisis.

These three chapters show how the use of quantitative methods can be applied to different types of financial data. The originality of the contribution of the entire thesis is therefore twofold. Firstly, an interdisciplinary production field was examined and, secondly, a workflow for the analysis of financial statements data was outlined. The interdisciplinary approach has allowed to explore the domain of informometrics in accounting and finance, offering a more complete and comprehensive understanding of the analysis of financial documents and paving the way for improvements in the field. The workflow proposed in this study allows to create an application for quantitative research, integrating both textual and numerical analytical methods, which could further improve the analysis of financial documents. Informetrics methods, i.e. the joint use of data science, computer science and accounting and finance knowledge have been applied to generate the tool which, following a systematic workflow, will allow the joint analysis of textual and numerical data in accounting and financial documents.

This study makes a valuable contribution to the field of financial document analysis by presenting a new and innovative approach to information metrics in accounting and finance. The interdisciplinary approach, systematic protocol and quantitative research application presented in the study can help improve the quality and accuracy of financial analysis and pave the way for further research in this area.

5.2 Theoretical and Practical implications

This has important theoretical as well as practical implications. In theoretical terms, the study emphasizes the importance of an interdisciplinary approach to financial document analysis, which can provide a more comprehensive and holistic understanding of financial document analysis. Second, the proposed workflow for the analysis of financial statements can serve as a theoretical framework for future research in this field. Third, the study highlights the importance of incorporating both textual and numerical analytical methods in financial analysis, which can improve the

accuracy and objectivity of the analysis. In practical terms, the study provides practical guidance for accounting and finance professionals to analyze financial statements more accurately and objectively. Secondly, the proposed application for the analysis of financial documents can be useful to various public/private sectors, as well as financial analysts and investors to evaluate the financial performance of companies. Finally, the interdisciplinary approach used in this study can help bridge the gap between accounting and finance and information technology. This makes financial analysis more efficient and less expensive, making it accessible to a wider range of investors and financial institutions.

5.3 Future agenda

There are many future research opportunities related to financial analysis. First, there is a need to develop data mining and data warehousing techniques to integrate unstructured data (such as texts) with structured information (such as numbers). The integration of unstructured and structured data will provide a more complete and accurate understanding of financial phenomena and support informed decisions in the financial world. Second, one must also keep in mind that the use of advanced technologies also presents certain risks, such as the possibility of developing models based on incomplete or erroneous data; therefore, there is a need to develop ad hoc techniques and methodologies to make textual analysis in accounting and finance take on greater consistency. This can be done through careful evaluation of the accuracy and robustness of artificial intelligence and machine learning models used in financial analysis. Finally, there is a growing need to develop methodologies to ensure transparency in the use of advanced technologies in financial analysis. This includes the need to make models more interpretable so that their results can be easily understood and trusted.

Table 5 Metods and Implications covered in the thesis

Chapter	Design	Search	Appraisal	Analysis	Synthesis	Implications
A bibliometric review of text mining in accounting and financial research	**RQ1)** How has the use of text mining in economics and finance evolved? **RQ2)** What are the themes that promote the use of text mining in A&F? **RQ3)** What are the themes that promote the use of text mining in management?	Database: WoS Query: "text* mining" OR "text* analysis" From: SCI-EXPANDED, SSCI Timespan: 2011-2022 Language: English	• Refined by Research Area:Business OR Business Finance OR Management • Refined by Document Types: Article and Proceedings Paper • Sample 1,280 publications	• Impact Metrics • Bibliometric analysis • Conceptual structure: Thematic evolution	High-impact A&F publications relate to non-financial reporting while low-impact publications relate to accounting information High-impact management publications concern bibliometrics, while low-impact publications concern consumer knowledge **Visual outcomes**: Strategy diagrams	Understanding the spread and adoption of text mining methods in accounting and finance and management disciplines, it is possible to identify the opportunities and limitations of these methods and develop a deeper understanding of their potential and uses in different contexts. This can help develop a deeper understanding of the factors that influence the effectiveness and efficiency of these methods and guide their further development and application.
Financial Configurations of Italian private hospitals: an evolutionary analysis	**RQ1)** What financial aspects best explain the heterogeneity between private hospitals? **RQ2)** What are the configurations of private hospitals? **RQ3)** How have these configurations changed over time?	Database: Aida Query: private hospitals and nursing homes Timespan: 2008 – 2012 - 2016	• Drop out hospitals that do not have financial autonomy, legal personality, and profit-making purposes • Drop out hospitals that do not provide ordinary or day hospitalization services • Sample 198 Italian accredited private hospitals	• PTA • Hierarchical clustering • Configurations approach	Four main financial configurations are identified, three of which worsened during the macroeconomic crisis of 2008-2016 Italian private hospitals have unsustainable business models and are highly exposed to banks **Visual outcomes:** Strategic Diagrams	Defining the essential features of Private Hospitals and cataloguing financial features represent the first empirical challenge of research that seeks to provide insightful suggestions to practice and research (especially in strategy and governance fields). Configurational studies are a promising development; they require appropriate research methods for their potential to be realized. Clustering methods are central to this avenue of inquiry.
Emotional Language and Financial Performance: Evidence from CEO Discourse during the Covid-19 Pandemic	**RQ1)** What communication strategies have companies adopted to communicate their financial results to their stakeholders? **RQ2)** How do CEOs use emotion and sentiment language to communicate their companies' financial results? **RQ3)** Is there a correlation between the affective and emotional values of letters to shareholders and total return?	Database: Standard and Poor's 500, Euro Stoxx and Nikkei indexes e Bloomberg Query: letter to the shareholders from the CEO or the BOD Timespan: 2019 - 2020	• Including all the companies listed in the Standard and Poor's 500, Euro Stoxx and Nikkei indexes • Including the text of the CEO's or BoD's letter to shareholders, headquarters, and sector • Including Annual Total Return: dividends, capital gains, payouts, distributions, and other forms of investment return • Sample 998 companies	• Topic Modeling • Emotional and Sentiment Analysis • Pearson's correlation	12 sectors, with a predominance of the industrial and financial sectors. U.S. companies were more adversely affected by the pandemic than companies in other countries in terms of performance. Letters from CEOs contained a great deal of information written clearly and concisely. 14 main topics, including corporate governance, economic performance, sustainability, digital transformation, and customer relations. CEOs focused on building trust and conveying optimism for the future while facing significant challenges, as indicated by the analysis of emotions and sentiments and their correlation. **Visual outcomes:** Table and Charts	Analyzing textual information in financial documents provides new insights for financial forecasting and auditing. The COVID-19 pandemic has accelerated the transformation of the sector and companies must adapt to remain competitive. Policy makers should encourage restructuring of the sector with funding for digitization and sustainability. Businesses too need to invest in digital technologies and remote work platforms to ensure business continuity in future crises and maintain a strong customer orientation by adapting to their changing needs.