UNIVERSITY OF PAVIA

DOCTORAL THESIS

The role of textual data in finance: methodological issues and empirical evidence

Author:
Roberta SCARAMOZZINO

Supervisor:
Prof. Paola CERCHIELLO

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

XXXIV CYCLE
Electronics, Computer Science and Electrical Engineering

2022

"Learning is experience. Everything else is just information."

ALBERT EINSTEIN

"Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less."

MARIE CURIE

UNIVERSITY OF PAVIA

Abstract

Electronics, Computer Science and Electrical Engineering

Doctor of Philosophy

The role of textual data in finance: methodological issues and empirical evidence

by Roberta SCARAMOZZINO

This thesis investigates the role of textual data in the financial field.

Textual data fall into the more extensive category of alternative data. These types of data, such as reviews, blog post, tweet, are constantly growing, and this reinforces the importance in several domains.

The thesis explores different applications of textual data in finance to highlight how it is possible to use this type of data and how this implementation can add value to financial analysis. The first application concerns the use of a lexicon-based approach in the credit scoring model. The second application proposes a causality detection between financial and sentiment data using an information-theoretic measure, the transfer entropy. The last application concerns the use of sentiment analysis in a network model, called BGVAR, to analyze the financial impact of the Covid-19 Pandemic.

Overall, this thesis shows that combining textual data with traditional financial data can lead to a more insightful knowledge and, therefore, to a more in-depth analysis, allowing for a broader understanding of economic events and financial relationships among economic entities of any kind.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
AIS	**A**rtificial **I**mmune **S**ystems
ANNs	**A**rtificial **N**eural **N**etworks
AUC	**A**rea **U**nder **C**urve
BGVAR	**B**ayesian **G**raphical **V**ector **A**uto **R**egressive
CDS	**C**redit **D**efault **S**wap
DJIA	**D**ow **J**ones **I**ndustrial **A**verage
EDA	**E**xploratory **D**ata **A**nalysis
eWOM	**e**lectronic **W**ord of **M**outh
FN	**F**alse **N**egative
FP	**F**alse **P**ositive
FSA	**F**inancial **S**entiment **A**nalysis
GAs	**G**enetic **A**lgorithms
IE	**I**nformation **E**xtraction
iid	**i**ndependent **i**dentically **d**istributed
IR	**I**nformation **R**etrieval
KDD	**K**nowledge **D**iscovery in **D**atabases
KNN	**K** **N**earest **N**eighbor
KW	**K**ruskal **W**allis
MB	**M**ega **B**yte
MN	**M**arkov **N**etwork
NLP	**N**atural **L**anguage **P**rocessing
ML	**M**achine **L**earning
POS	**P**art **O**f **S**peech
S&P	**S**tandard and **P**oor
SMEs	**S**mall **M**edium **E**nterprises
SVM	**S**upport **V**ector **M**achine
TE	**T**ransfer **E**ntropy
tf-idf	**t**erm **f**requency **i**nverse **d**ocument **f**requency
TP	**T**rue **P**ositive
TN	**T**rue **N**egative
US	**U**nited **S**tates
VAR	**V**ector **A**uto **R**egressive

Dedicated to my family.

Chapter 1

Introduction

"Without data you're just another person with an opinion."

W. Edwards Deming

1.1 Alternative data in Finance

The new era of digital world is characterized by an availability of data never seen before. Among these data, the so-called alternative data stand out. By alternative data we mean a type of unstructured data, for example a blog post, a social media or e-commerce content. These data are used in several domains such as health, technology, politics, communication, and also finance sector. In the latter domain, alternative data are used in different fields of the financial sector from risk management to equity selection, portfolio diversification and security.

The goal of using alternative data is to make sense of this unstructured data in order to improve the accuracy of decision process and obtain then a competitive advantage, considering that, most of the data that is daily created is unstructured.

According to a report by Grand View Research[1], the market value of alternative data was estimated to be 1.06 billion dollars in 2019 and growing to a forecast of 17.3 billion in 2027, with a composite annual growth rate of 40% from 2020 to 2027. The greatest demand for this data is coming from hedge funds, private equity, life insurance companies, banks and fintech. The financial sector is increasingly investing resources in order to achieve a competitive advantage through the use of this data knowing that information is one of the most valuable asset for a business.

In the financial domain, using alternative data has led to a creation of more advanced ways of granting credit, especially in less developed countries as those in Africa or in South America. In those countries, it is very difficult to verify and search for a solid credit history for loan applicants. Using unstructured data is helping financial institutions to solve this problem which, however, does not only concern underdeveloped countries. In more developed countries such as the United States, new studies have been made to understand the benefits of using alternative data for both individuals and the entire economy. A recent study (Feinstein, 2013) shows that 41% of minority consumers cannot be classified using normal credit scoring models. On the other hand, 81% of this percentage can be evaluated using alternative data. This minority consumers are the so called "credit invisible" and they are mainly represented by students, recent graduates, immigrants or all minorities that are often rejected because of the lack of information about them.

[1] https://www.grandviewresearch.com/industry-analysis/alternative-data-market

In chapter 3 we deal with an implementation of a classic credit scoring model using textual data.

As said before, using alternative data in the credit scoring model is a relatively recent strategy, while the analysis of texts from blogs, tweets, forums in the financial field has been used for longer. In 1966 Stone, Dunphy, and Smith, 1966 described how contents are evidence of the behavior of individuals' activities also in terms of economic and financial behavior. A detailed sentiment analysis research, ranking existing studies based on approaches and applications, was conducted by Ravi and Ravi, 2015 while a survey of techniques and the related fields was conducted by Medhat, Hassan, and Korashy, 2014.

Sentiment analysis uses textual data for investigating opinions and thoughts about a given topic. With the evolution of social media and the web 2.0, people have increasingly shared their emotions, ideas, thoughts, and opinions. The Covid-19 crisis has forced the whole world to reduce, and in most cases eliminate social interactions. Daily events and activities that previously took place offline had to move online, increasing even further this phenomenon. In the marketing field, attention to consumer opinion obtained from the web is extremely important. In this framework, the electronic Word of Mouth (eWOM) is defined as *"any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet"* (Hennig-Thurau et al., 2004).

However, this concept is no longer linked exclusively to marketing but to any type of activity or service that aims to satisfy customers. An example is represented by a financial operator who wants to make better decisions for its customer's investment portfolio or borrowers to be able to lend safely more money to lenders. In order to succeed in such tasks, information on sentiment is needed, in order to make companies be able to take more accurate decisions. Financial and economic news are constantly affecting share prices, therefore many traders and investors are including the results of their sentiment analysis application to empower their decision making process. The first problem encountered in the extraction of opinion from financial news is that it requires a specific language. To overcome this problem, Loughran and McDonald, 2011 created a lexicon of positive, negative, and neutral words in finance.

Mishev et al., 2020 presented a detailed overview of NLP-based sentiment analysis approaches in finance, and Man, Luo, and Lin, 2019 conducted a survey of financial sentiment analysis (FSA). Kearney and Liu, 2014 conducted a survey on textual analysis in finance, focusing on the sources of information, the methods of analysis, and financial models applied to examine whether the addition of textual analysis had a positive influence. Another study conducted by Smailović et al., 2013 analyzed whether the sentiment expressed through tweets can be used to predict share prices. Tweets were classified through support vector machine classification and then the Granger's causality test was used. The polarity obtained from these results was able to predict the movement of share prices a few days in advance. All this research shows the positive impact of sentiment analysis on stock returns and trading volumes. The studies conducted on this subject are very extensive and we will mention more in the next chapters.

Another important factor to consider when it comes to financial sentiment is globalization. Companies are closely connected, therefore it is important to know opinions about a company, not only to understand its share price fluctuations but also the effect on its business network.

At the financial level, a broader analysis that considers the interactions between the different players is widely studied at the banking level, especially in the field of systemic risk. A survey of systemic risk was conducted by De Bandt and Hartmann, 2000.

Systemic risk is the risk of collapse of an entire financial system starting from an initial shock (trigger event) such as the 'rush to the counter', problems with payments from a financial institution, failure of a large bank which then spreads to the rest of the market creating instability in the entire sector. In a recent review article Caccioli, Barucca, and Kobayashi, 2018 summarize the modeling of financial systemic risk.

However, since the financial crisis of 2008 -2009, it has been understood that it is no longer possible to consider a single sector, such as banking, in isolation. On the contrary, all the actors of the economy must be considered. A crisis born in American sub-prime mortgages was able to propagate and heavily impact all the economies in the rest of the world. The current crisis due to the pandemic has highlighted the same issues and has demonstrated that nowdays it is no longer possible to consider an event in isolation. On this basis, and also for the growing development of social media and in general sharing through the web, in today's world, it is essential to understand the interactions between all the actors of the economic system. First of all to understand how an event spreads and also to understand its intensity and repercussions that may also exist. The interactions between companies and also across sectors using textual analysis are considered in chapter 4 and 5.

Over the past decade, following the 2008 crisis, the literature on network-based systemic risk has grown considerably. Kou et al., 2019 examined the latest studies and researches on financial systemic risk with machine learning methods such as big data analysis, network analysis, and sentiment analysis. Acemoglu et al., 2012 analyzed the interconnections between different companies and sectors and showing how these generate "cascading effects" such that a shock in a production sector spreads directly to customers of that sector and then to the rest of the economy . In subsequent work, Acemoglu, Ozdaglar, and Tahbaz-Salehi, 2015a demonstrated how the propagation mechanism depends on the extent of a shock. When a shock is above a certain threshold, it has a negative impact on the financial structure making it more fragile and facilitating the spread of the shock to other institutions.

The application of sentiment in the network model has also increased in the last years. Wan et al., 2021 applied NLP techniques to understand the sentiment, over 7 years, of 87 companies by investigating the spread among all companies and evaluating the variation in prices and volatility . Cerchiello and Giudici, 2016b proposed a systemic risk model based on big data using two types of information: the price information and textual data from financial tweets. This information has been combined using a Bayesian approach. Recently Nyman, Kapadia, and Tuckett, 2021 analyzed textual data on the financial market for identifying metrics capable of capturing developments in the financial system. The study showed how these metrics are able to capture the opinions and sentiments of a significant financial events. The results showed how the negative sentiment provided from the analysis was followed in mid-2007 by the collapse of the Lehman brothers.

The analysis of the relationships between different sectors of the global economy through textual analysis, specifically sentiment analysis, is conducted in Chapters 4 and 5.

1.2 Mission

This thesis aims to apply the fundamental concepts of data science to the financial sector with particular regard to the use of alternative data, namely textual data.

The main goal is to demonstrate how using textual data can add value to financial analysis. We will specifically look at different applications in this area, evaluating their advantages and disadvantages. We apply machine learning techniques combining classical data and alternative data and highlight how alternative data should not replace the former but enrich them.

1.3 Structure

The thesis is structured as follows:

The first chapter introduces the main topics of the thesis and explains its structure.

Chapter 2 defines the theoretical background of the main fields discussed later. The chapter is divided into five sections:

(i) data science

(ii) textual analysis

(iii) credit scoring

(iv) information theory

(v) network analysis.

In these sections we focus on the concepts that most help the understanding of the rest of the manuscript.

Chapter 3 proposes a credit scoring model in which textual data are added to the typical financial data relating to bank transactions. The methodology used to manage textual data is a lexicon-based approach, specifically dictionary-based. Text mining is applied to bank transactions and classifies them into macro classes. These macro-classes are then used to create two variables (frequency and the total amount spent by the client for the specific category) and inserted in a credit scoring model together with the typical financial variables (current account balance, the sum of income, the total number of movements, and so on). The final aim is to verify whether such text-based categories can act as effective predictors of bank default. The important result obtained through our analysis is determined by the distribution of errors. Through the application of textual classification the type I error or False positive decreases. The costs associated with type I errors are higher than type II errors. This is because, in this field, accept false positives means accept those who will not pay. This generates more losses than reject those who would have paid. Type 1 are actual cost, type 2 are opportunity for revenues lost.

This work fits into the literature on textual analysis for credit scoring. The main innovation of this analysis is the source of textual data, not social data (as usually used in this area) but bank transactions.

Chapter 4 proposes a causality detection between financial and sentiment data based on an information-theoretic measure, analyzing the top 50 companies in the Standard & Poor (S&P) index during the period from November 2018 to November 2020. Considering the capitalization of the companies studied in our sample, analyzing the individual connections could be reductive, so we decided to focus on a

macro level highlighting the influences between the various sectors. In the analysis made, we observe how most of the connections occur by considering the same source (from price variables to price variables and from sentiment variables to sentiment variables) but the connections between the different sources are also relevant. We note how the influence between the different companies goes beyond the sectors, indeed we can say that the influence between different sectors is more relevant.

This work fits into the literature on textual analysis in finance. The main innovation is the use of an information theoretic measure, called entropy, considering opinions and stock market data.

Chapter 5 proposes a network based evidence of the financial impact of the Covid-19 pandemic through a model that combines two types of information i.e. sentiment data and market prices. The goal is to highlight the connections among companies considering the period before and during the pandemic crisis. We demonstrate that the crises resulting from the COVID-19 pandemic affects all US production systems, suggesting that although the U.S. market might have appeared resilient resilient during the first wave, it certainly did not appear so during the second wave. We also observe that the response to the pandemic is linked to the business where companies are operating. The healthcare sector, as expected, is the most affected while the financial sector is the most stable. We also observed how the outbreak of the pandemic created a greater networks density which suggests that a widespread shock produces more interconnections and this determines a greater vulnerability in terms of systemic risk.

This work fits into the literature on impact of Covid-19 pandemic through textual analysis. The main innovation is the use of an advanced network model combining financial information and sentiment data.

Chapter 6 presents the conclusion of this work pointing out the advantages of using textual data in finance.

Chapter 2

Litterature review

"The goal is to turn data into information and information into insight."

Carly Fiorina

This chapter introduces the key notions discussed in the thesis and their related literature. It is structured in five sections:

(i) data science

(ii) textual analysis

(iii) credit scoring

(iv) information theory

(v) network analysis.

Each section is composed by an introduction to the argument, an explanation of the basic concepts, and insights into the most pertinent aspects of those fields associated with the following chapters.

The goal is to give the reader a vision, as far as possible, complete with all the topics covered in this thesis.

2.1 Data Science

In the era we are living in, the highest value available to companies is data. Research by the Domo[1] company estimated that, in 2020, every person in the world created 1.7 MB of data every second. Online life is increasingly predominant and the amount of data is destined to grow exponentially. It is estimated[2] that in 2025 this value will rise to 463 exabytes[3]of data per day per person. Considering this growth, the importance of being able to read and interpret data in any field of application, from finance to marketing to healthcare, becomes central.

The science of extracting knowledge from data is known as data science. It is a transversal discipline, which includes statistics, computer science, mathematics, and also more managerial domains. As the name suggests, the essential part of this discipline is data, which must be collected and processed and then read and interpreted in order to extract value.

Six stages of the data science life cycle can be identified by several authors:

[1] https://www.domo.com/

[2] https://www.raconteur.net/infographics/a-day-in-data/

[3] $1 exabytes = 1 trillionbyte = 10^{18} byte$

- Problem definition

  Before collecting any data, it is essential to understand what is problem we want to solve. The choice of data to be used is strictly linked to this.

- Data collection

  Once the problem has been defined, the most pertinent data can be collected, considering that not all data are useful for all problems. Two macro classes of data can be defined: structured and unstructured data. The first type is data with a precise format that are stored in databases, the second is data without a defined structure.

- Data preparation

  This step includes data cleaning, management of missing data, correction of incorrect values, elimination of duplicates, and structuring of data for the algorithm. It is usually the most time-consuming phase.

- Exploratory Data Analysis (EDA)

  This phase concerns the study of the data key characteristics. There are several techniques used to understand the dataset such as extraction of important variables, identification of outliers, identification of relationships between variables and anomalies.

- Model Building

  Selection and design of the model to be applied to solve the problem. The model building is itself divided into sub-steps.

  Algorithm selection, this depends on the available data and the objective to be pursued. We identify two broad areas of areas of learning: supervised learning and unsupervised learning.

  Training, after the algorithm has been chosen, it is trained on data (training data). In this step the algorithm learn rules from data that can be apply to future unknown data.

  Testing, is the forecasting phase where new data (test data) are given to the model. The model is evaluated on how it classifies/processes the test set on the basis of what it has learned in the training phase.

- Results and Communication

  The last step of this cycle is the analysis of the results obtained and the relative communication usually through graphs, reports, and tables. This phase is closely linked to the first, which is why we talk about cyclicality.

Before going on, it is necessary to define some notions that are often confused in this field. For example, the term *data science* is often used in place of other terms like *machine learning*, *artificial intelligence*, and *deep learning*, but these terms are not equivalent.

Artificial intelligence (AI) is the discipline that deals with the creation of machines (hardware and software) capable of solving problems and performing tasks typical of the human intellectual sphere, thus simulating cognitive abilities. (Müller and Bostrom, 2016)

Machine learning (ML) refers to algorithms based on mathematical and computational approaches that can learn from data and then apply the learned rules to new

data, and whose accuracy improves over time (based on the given examples) (Qiu et al., 2016). The learning techniques can be supervised or unsupervised. In the first case, we have a priori knowledge of data labels. The machine is then given both inputs and outputs and the goal is to find a relationship between these (a function that binds inputs and outputs). The main tasks of this type of learning are classification and regression. In the second case, we do not know the data labels. The machine will then have the objective of making inferences on the structure of the data. The outputs are not available. The main tasks of this type of learning are clustering, association, and dimensionality reduction. Alloghani et al., 2020 in their systematic review of academic papers from 2015 to 2018 show that in the first case the most used algorithms are decision tree, support vector machine, and Naïve Bayes, while in the second case k-means, hierarchical clustering, and principal component analysis.

Deep learning is a form of machine learning where artificial neural networks are used. Artificial neural networks are algorithms that mimic the structure, functioning, and connections of human neurons. Initial data is processed in numerous interlinked steps (Zhang et al., 2018).This form is especially used when a large quantity of data is provided with a particularly high number of features.

As can be seen in figure 2.1, deep learning is part of machine learning which is a branch of artificial intelligence, while data science is involved in these three areas (in addition to those already mentioned previously).

(Kulin et al., 2021)

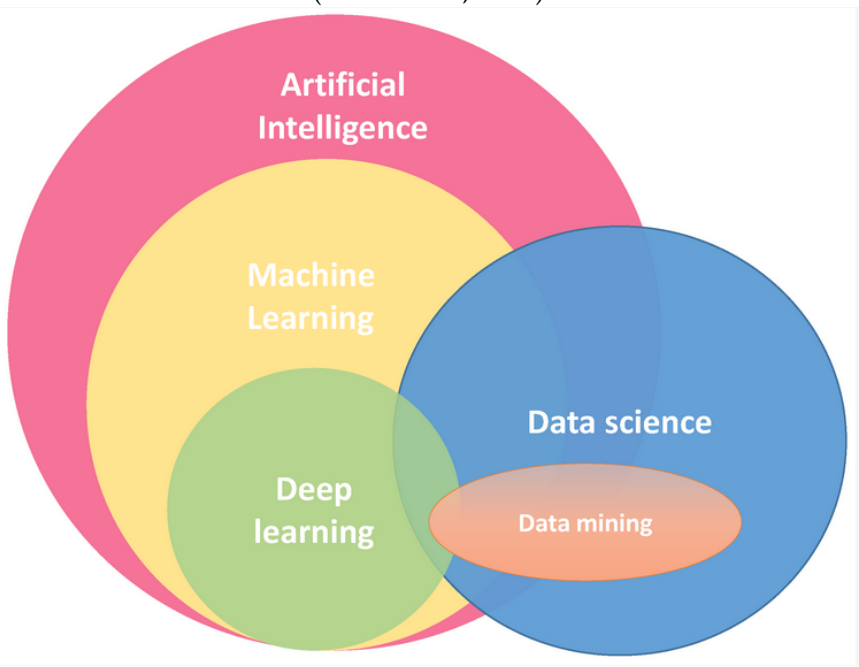


FIGURE 2.1: Data science, artificial intelligence, machine learning and deep learning

2.1.1 Big Data

As mentioned above, the fundamental element of data science is data. The more data are available, the more accurate the knowledge obtained through their analysis is. The great potential of big data is explained by Mayer and Cukier: *"the idea is that we can learn from a large body of information things that we could not comprehend when we used only smaller amounts."* (Mayer-Schoenberger and Cukier, 2013)

When speaking about Big Data, the 3 Vs must mentioned: volume, velocity and variety (Laney et al., 2001).

- Volume: a large amount of data.
- Velocity: these data are collected in a very short period.
- Variety: the data are increasingly heterogeneous, (i.e. they come from a multitude of different sources).

However, the definition of Big Data has recently expanded to include two other features: veracity and value. For this reason, so we talk about 5v.

- Veracity: the data must be truthful, reliable.
- Value: it is not enough to have data, value needs to be created from them.

Types of data

Three large data families are generally defined: Structured, Semi-structured and Unstructured.
Structured, also called quantitative data, are data whose type (date, amount, geolocation) can be precisely defined. They are organized according to rigid schemes and tables (Flach and Lachiche, 2004).
Semi-Structured, are data with a structure but cannot be organized in standard tables.
Unstructured also called qualitative data, they don't have a default template (for example text, social media activity, video files) (Eberendu et al., 2016).

The former is the most widely used, both for the fact that they are very frequent but above all for their ease of use. Despite this, the latter type is becoming more used because, at the expense of a more difficult cleaning, there are some advantages in terms of performance. [4]

Jeff Bezos, founder and president of Amazon declared his opinion about the "superiority" of the latter. *"The thing I have noticed is when the anecdotes and the data disagree, the anecdotes are usually right. There's something wrong with the way you are measuring it."*

2.2 Text Mining

Closely related to the concept of data science is that of data mining.

"At a high level, data science is a set of fundamental principles that support and guide the principled extraction of information and knowledge from data. Possibly the most closely related concept to data science is data mining—the actual extraction of knowledge from data via technologies that incorporate these principles." (Provost and Fawcett, 2013).

Within the broad field of data science is data mining. With data mining, we define the activities of extracting information from a large amount of data with the aim of discovering significant facts.

The words data mining and knowledge discovery in databases (KDD) are often interchanged but they do not coincide in practice. The former is the entire process of searching for new knowledge. Data mining is the application of specific algorithms to extract patterns. It represents a specific step of the entire process.

[4] https://www.ibm.com/cloud/blog/structured-vs-unstructured-data

When textual data is used for researches in data mining we generally can talk about text mining (Dörre, Gerstl, and Seiffert, 1999).

"In text mining the patterns are extracted from natural language text rather than from structured databases of facts" (Gupta, Lehal, et al., 2009).

The first researchers to talk about obtaining information from textual data were Feldman and Dagan, 1995, introducing the concept of knowledge discovery from text (KDT) . A clear survey of text mining is carried out by Allahyari et al., 2017.

2.2.1 Techniques

- Information Extraction (IE)

  It is the technique of automatic extrapolation of specific information relating to a topic from textual data sources (Cowie and Lehnert, 1996).

- Information Retrieval (IR)

  A survey in information retrieval is done by Faloutsos and Oard, 1998. This technique retrieves relevant information from systems. The aim is to satisfy the information need of the user. When a query is carried out, the objective of the IR is to respond accurately, selecting the one most relevant to the user's request from the documents. The goal is to find information rather than understanding it.

- Data Mining

  Data mining is the set of techniques and methodologies for extracting useful information from large amounts of data. From the viewpoint of data mining, Wu et al., 2013 suggest a big data analysis model .

- Natural Language Processing (NLP)

  NLP is an interdisciplinary research field that aims to develop models capable of understanding and analyzing natural language, both written and spoken. It deals with analyzing the syntactic structure of texts and also the semantic one. Manning and Schutze, 1999 wrote the first book on statistical natural language processing.

- Categorization

  Text Categorization, or text classification, aims to assign natural language documents to a set of predefined classes.

2.2.2 Step

In figure 2.2 it is presented the text mining process flow. The steps in text mining are:

1. Data collection

   The first phase is the collection of the data that are relevant for the purpose of the analysis.

2. Text preprocessing

   It concerns the cleaning of data and the creation of variables useful for analysis, starting from the raw data.

3. Text transformation

   Also known as attribute generation, it is the representation of the document through two approaches: bag of words and vector space. Bag of words is a representation of the text that does not consider any information about the order and the structure of the words in the document but considers only how many times they appear. Vector space is an algebraic model where documents are represented as vectors of identifiers.

4. Feature selection

   It is the selection of variables considered relevant for the objective, eliminating the redundant ones or the ones which do not add value to the analysis. It allows for simpler and faster subsequent analysis.

5. Data mining methods

   At this point text mining and data mining overlap; data mining techniques are used to analyze the data obtained in the previous stages.

6. Interpretation/Evaluation

   It is the evaluation of the results obtained.

Source: (Segall, Zhang, and Cao, 2009)

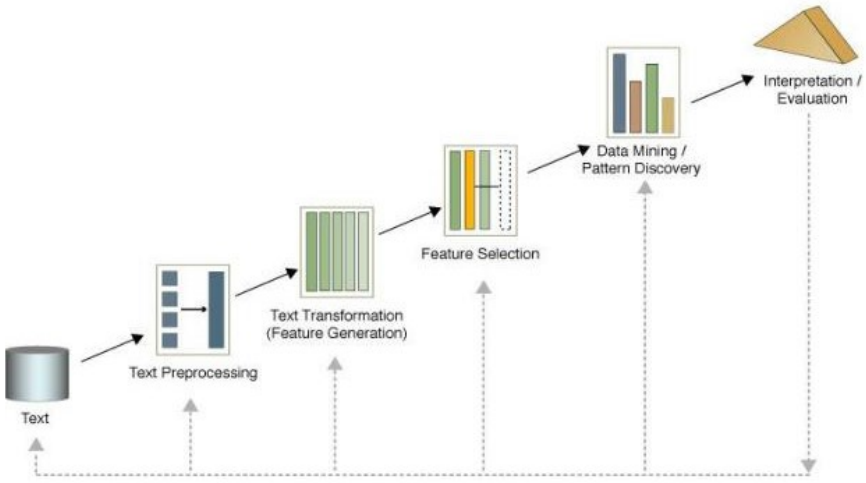


FIGURE 2.2: Text mining process flow

A key role of the process is step two. This is confirmed by Uysal and Gunal, 2014. They noticed that the choices made in this step can improve the accuracy of the classification.

In detail, text preprocessing concerns some specific tasks such as:

- Tokenization

Tokenization is usually the first step of pre-processing. In this task, the entire document is divided into pieces (token) and some characters such as punctuation marks and accents are removed.

- Filtering

  Filtering is a step regarding the removal of words that are not significant for the document. One of the filtering processes more used is stop-words removal. The stop words are words that are frequent in documents but not relevant, for example, articles, prepositions, conjunctions, HTML for web pages and so on.

- Lemmatization

  Lemmatization is a way to associate to a unique lexical item (the lemma) the inflected forms. It is base on the inflectional morphology rules of the concerned language and it is necessary to specify the part-of-speech (POS) of each word. The possibility of error and the significant amount of time necessary to carry out lemmatization, make this step less preferred compared to stemming.

- Stemming

  Stemming is a technique that aims to get the root (stem) of each word with the objective of reducing inflectional forms. For example, *connected, connecting, connections, connection* are all turned into *connect* while *looked, looking, looks* become *look*. It is a technique that depends on the language. Porters Stemming is one of the most used for English language. Jivani et al., 2011 presented a survey of different algorithms .

2.2.3 Text Classification

Text classification can be done in two ways: manually or automatically. In the first case, a person analyzes a set of documents and classifies them on the basis of their content. This type of classification is often more accurate, but the major problem is that it is time-consuming (Nigam et al., 2000).

In the latter case, automatic techniques are applied to classify documents. In this case, the process is faster but errors are more likely.

Automatic classification can be performed using several approaches (Dalal and Zaveri, 2011):

- Rule-Based

  Typically it involves the use of dictionaries. In dictionary the most representative keywords for of each category are provided, and on this basis the machine analyzes the entire text and assigns it to a given class.

- Machine Learning algorithm

  In this case, the machine learns rules starting from past observations during the training phase, where documents are provided with labels depending on the type of classification. The most used example of machine learning algorithms are k-means, naive bayes classifier, k-nearest neighbor (KNN), support vector machines (SVM). Machine learning represents a solution that assures a fast completion of the classification task.

- Hybrid systems

Hybrid systems are combinations of the two previous approaches. The advantage of hybrid methods is that they make more accurate machine learning models completing them with specific rules for the most "problematic" cases (Ghiassi, Skinner, and Zimbra, 2013).

Sentiment Analysis

Sentiment analysis is one of the most well-known applications of text classification and is widely used especially in the marketing field. It is also known as opinion mining. The aim of sentiment analysis is to detect and extract opinions from the text. It has become an increasingly relevant topic given the growth of unstructured data, such as opinions and reviews on the web.

Sentiment analysis uses techniques from Natural Language Processing, Information Retrieval, and Data Mining (Ravi and Ravi, 2015).

These techniques are applied to a large datasets from different sources such as social networks, blogs, newspapers, reviews. After these texts are pre-processed, the opinions in them are determined through the study of the words and punctuation that compose the text. Particular attention is given to the context in which these texts are inscribed. The main focus of sentiment analysis is the polarity (whether the opinion is positive, negative, or neutral) but also emotions are addressed (calm, happiness, sadness). The most used data sources are social media platforms. In particular, Facebook is the most used because of the amount of data generated, and Twitter due to the maximum number of characters of the tweets and the emoticons that facilitate the sentiment identification.

An analysis of the state of art of sentiment analysis techniques and applications was recently conducted by Medhat, Hassan, and Korashy, 2014, while comparative study of these techniques was conducted by Devika, Sunitha, and Ganesh, 2016 and Aggarwal and Zhai, 2012 wrote a complete book. Ravi and Ravi, 2015 reviewed research in sentiment analysis, from 2002–2014, classifying studies based on opinion mining tasks, approaches, and applications and listed available public datasets on sentiment analysis.

Some problems related to sentiment analysis were highlighted in some researches. For example, Tang, Tan, and Cheng, 2009 in a survey on sentiment detection discuss four problems: subjectivity classification, word sentiment classification, document sentiment classification based on machine learning techniques, and opinion extraction problem . On the other hand, Feldman, 2013a focused on five problems: document-level sentiment analysis (the author's opinion), sentence-level sentiment analysis (single opinion in each sentence), aspect-based sentiment analysis (different opinion for each aspect of an entity), comparative sentiment analysis (presence of comparable opinions instead), and sentiment lexicon acquisition (data acqusition).

An important part of sentiment analysis is feature extraction. Some important features are:

- *Term frequency–inverse document frequency (tf-idf)*: it is a function used to measure the importance of a term. In general, the most important terms are those present in the document but less frequent.
- *Part of Speech (POS)*: words classes, such as nouns, verbs, adjectives are indicators of opinion.
- *Negations*: negations change the meaning of a sentence, for example, add a not to the sentence "I liked " changes radically the polarity of the sentence.

- *Topic-oriented features*: often the topic of the text affects the real meaning of the sentiment.

Also, sentiment classification techniques can be divided into three categories (figure 2.3) as discussed on page 14 about text classification.

In rule-based approaches, two types of lists are used. The first one is the list of words associated with a positive polarity (like, good, beautiful, useful, interesting, love), the second one is composed by words associated with a negative polarity (bad, ugly, sad, boring, frustrated). In machine learning approaches, machine learning algorithm are applied in a training dataset to give machines examples of positively and negatively polarized texts in order to learn how to classify new texts.

Source: (Aqlan, Manjula, and Naik, 2019)

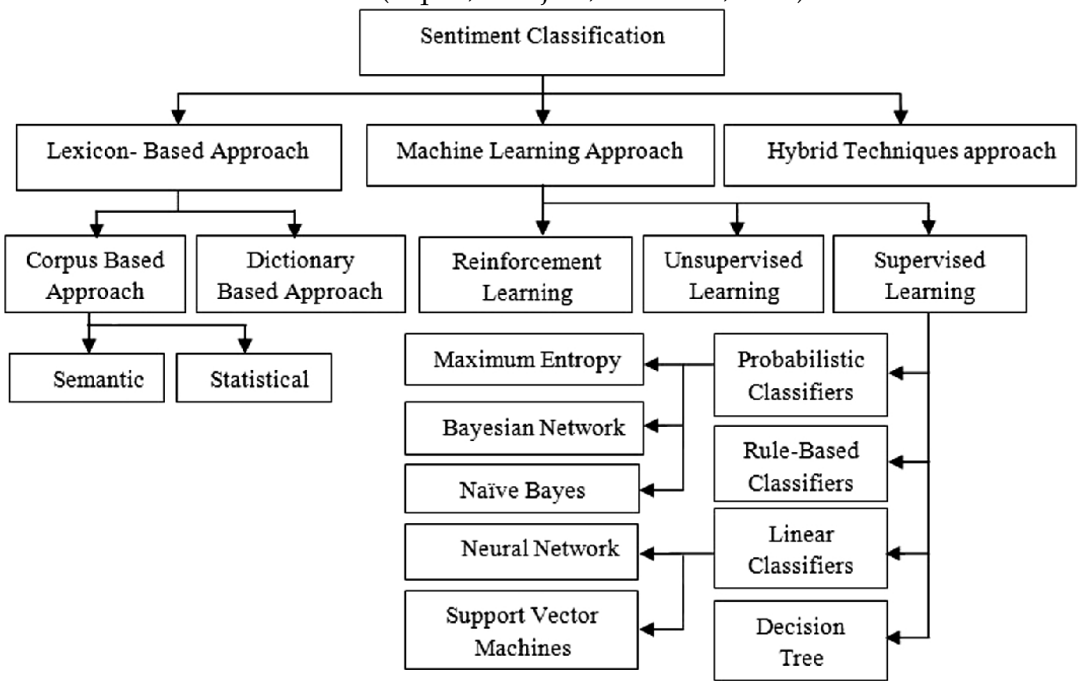


FIGURE 2.3: Sentiment Analysis techniques

2.3 Credit Scoring Models

Credit scoring model is a statistical analysis to estimate the probability of default.

"This credit scoring process will determine who should get credit and how much credit should be granted with the intention to minimize the risk of loan losses and delinquency rate due to the costly misclassifying error." (Eddy and Engku Abu Bakar, 2017)

It is often confused with the terms credit rating but they are different. Credit rating is the creditworthiness of a business or government and it is expressed through capital letter (AA, A, BBB, BB, B, CCC, CC, C, and D) while credit score is used for consumers and it is expressed through a number.

There are different credit score, the most used is Fico (Fair Isaac Corporation's credit scoring system). It is used by more than 90% of top lenders[5]. Fico score is composed by five components[6]: for 35% by Payment history, for the 30% by amounts owed, for the 15% by length of credit history, for 10% by new credit and for the last 10% by credit mix.

2.3.1 Credit Scoring Techniques

Credit scoring techniques are divided in two categories: statistical and machine learning (Mpofu and Mukosera, 2014).

Statistical-based credit scoring models

- *Linear regression* is a model that describes the relationship between a response variable (dependent variable) and one or more independent variables (Abdou and Pointon, 2011).
- *Discriminant analysis* is a technique that classify observations into groups based on explanatory variables (Khemais, Nesrine, Mohamed, et al., 2016).
- *Probit analysis* is a type of regression in which the linear combination of the independent variables is transformed into into its cumulative probability value from a normal distribution (Mpofu and Mukosera, 2014).
- *Decision trees* is a non-parametric method. It classify a variable based on other variables by constructing a tree structure in which we have root node, internal nodes, and leaf nodes (Song and Ying, 2015).
- *Logistic Regression* is a technique specific for binary classification problem. The difference with normal regression is the use of dichotomous dependent variable (Bolton et al., 2010).

Machine learning methods

- *Artificial Neural Networks (ANNs)* is a non-linear model with a structure similar to neurons in a biological brain. It is based on neurons and connection (synapses). Starting from a variable input we obtain classification based on pattern recognition capabilities (West, 2000).
- *Genetic Algorithms (GAs)* is an abstraction of Darwinian evolution. It replicates the natural selection process in which only the fittest individuals are selected from a population (Gordini, 2014).

[5] https://www.investopedia.com/terms/c/credit_scoring.asp

[6] https://www.myfico.com/credit-education/whats-in-your-credit-score

- *Artificial Immune Systems (AIS)* is a replication of the immune systems. The problems are solved through the features of memory,experience and perform pattern recognition (Kamalloo and Abadeh, 2010).

Marques, García, and Sánchez, 2013 review literature on the application of evolutionary computing to credit scoring

Abdou and Pointon, 2011 review 214 studies related to credit scoring and show how advanced techniques perform higher predictive ability than classical techniques.

In the categorization problem, like credit scoring, the most used metric for measuring the reliability of the techniques used are the confusion matrix and related metrics.

The confusion matrix is a table that calculates: true positives (TP) when the classifier assigns correctly the positive class, true negatives (TN) when the classifier assigns correctly the negative class, false positives (FP) when the classifier assigns incorrectly the positive class, and false negatives (FN) when the classifier assigns incorrectly the negative class.

The metrics *precision*, *recall(sensitivity)*, *specificity*, *accuracy* and *error rate* are based on confusion matrix. Precision indicates correct predictions, sensitivity is the ability of a classifier to find all positive instances, specificity is the ability of a classifier to find all negative instances. Accuracy is the ability of a classifier to correctly label instances. Finally, the error rate represents the rate of mistakes made in the classification process.

Mathematically speaking, these concepts are defined by the following formulas (Powers, 2020):

$$Precision = \frac{TP}{(TP + FP)}, \tag{2.1}$$

$$Sensitivity = \frac{TP}{(TP + FN)}, \tag{2.2}$$

$$Specificity = \frac{TN}{(TN + FP)}, \tag{2.3}$$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}, \tag{2.4}$$

$$Error\ Rate = \frac{(FP + FN)}{(TP + TN + FP + FN)}, \tag{2.5}$$

On the basis of these definitions, two further notions can be defined. The false positive rate and the false negative rate.

$$False\ Positive\ Rate\ (type\ I\ error) = 1 - specificity, \tag{2.6}$$

$$\textit{False Negative Rate (type II error)} = 1 - \textit{sensitivity}, \tag{2.7}$$

Finally, the ROC (receiver operating characteristic) curve is a graphical representation of the functional relation between sensitivity (definend in 2.2) and specificity (definend in 2.3) as the threshold value varies.

2.4 Information Theory

Information theory is a branch of computer science and telecommunications whose aim is to analyze and process phenomena relating to the calculation and transmitting of information (Guizzo, 2003).

The father of this discipline is Claude Shannon, an American electrical engineer and mathematician. The birth of information theory is considered to be the publication of his article "A mathematical theory of communication", in 1948, in the Bell System Technical Journal (Shannon, 1948).

Underlying information theory is the concept of entropy. Entropy is a measure that expresses the degree of disorder of a system, a concept born in thermodynamics but then applied in statistics. Shannon's intuition was to consider the lack of information in terms of disorder, as this lack of information causes uncertainty.

According to Shannon's first theorem (source coding theorem), it is impossible to compact a sequence of independent and identically distributed (iid) random variables of length that tends to infinity into a message shorter than their complete entropy without losing knowledge.

2.4.1 Basic concepts

Mutual Information

It is a metric for determining the reciprocal dependencies of two random variables which is formalized in (2.8).

$$I(X;Y) = \sum_{y \epsilon Y} \sum_{x \epsilon X} p(x,y) \log \left(\frac{p(x,y)}{p_1(x) p_2(y)} \right), \tag{2.8}$$

where $p(x,y)$ is the joint probability distribution function of X and Y, and $p_1(x)$ is the marginal probability distribution function of X and $p_2(y)$ is the marginal probability distribution functions of Y.

Measure the information that X and Y share.

Properties:

- if X,Y independent $\Rightarrow I(X;Y) = 0$, if X and Y are independent it follows that mutual information is 0;
- $I(X;Y) \geq 0$, mutual information cannot be negative;
- $I(X;Y) = I(Y;X)$, mutual information is symmetrical;

Entropy

The entropy is the average amount of data received with each message.

Starting from the concept of self-information, the information contained in an event x emitted by a source X is calculated as:

$$I(x) = -\log_b \mathbb{P}(x), \tag{2.9}$$

Where $\mathbb{P}(x)$ is the probability that the event x will happen.
The expected value of self-information is known as a source's entropy.
Mathematically:

$$\mathbb{H}(X) = \mathbb{E}[I(X)] = \mathbb{E}[-\log_b \mathbb{P}(x)], \tag{2.10}$$

If X is a discrete random variable the expected value is:

$$\mathbb{H}(X) = -\sum_{i=1}^{N} \mathbb{P}(x_i) \log_b \mathbb{P}(x_i), \tag{2.11}$$

If X is a continuous random variable the expected value is:

$$\mathbb{H}(X) = -\int \mathbb{P}(x) \log_b \mathbb{P}(x) dx, \tag{2.12}$$

Conditional Entropy

The conditional entropy is the amount of knowledge used to explain the value of one random variable X given the value of another random variable Y.

$$H(X|Y) = \sum_{k=0}^{K-1} H(X|Y = y_k) p(y_k), \tag{2.13}$$

Joint Entropy

The uncertainty associated with a number of random variables is measured by joint entropy.
Mathematically:

$$H(X;Y) = -\sum_{x}\sum_{y} P(x,y) \log_2[P(x,y)], \tag{2.14}$$

Properties

- $H(X;Y) > max[H(X), H(Y)]$

  The joint entropy of a set of variables is greater than or equal to all the individual entropies of the variables.

- $H(X;Y) \leq H(X) + H(Y)$

  The joint entropy of a set of variables is less than or equal to the sum of the individual entropies of the variables.

Relations
The mutual information can be expressed in terms of entropy:

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) = \\
&= H(Y) - H(Y|X) = \\
&= H(X) + H(Y) - H(X,Y) = \\
&= H(X,Y) - H(X|Y) - H(Y|X),
\end{aligned}
$$

where $H(X)$ is the marginal entropy of X and $H(Y)$ is the marginal entropy of X, $H(X|Y)$ and $H(Y|X)$ are the conditional entropies, and $H(X,Y)$ is the joint entropy of X and Y.

In the figure 2.4 we see a Venn diagram where the relationships between the different entropies defined previously are displayed.

Source: Deep dive into deep learning [7]

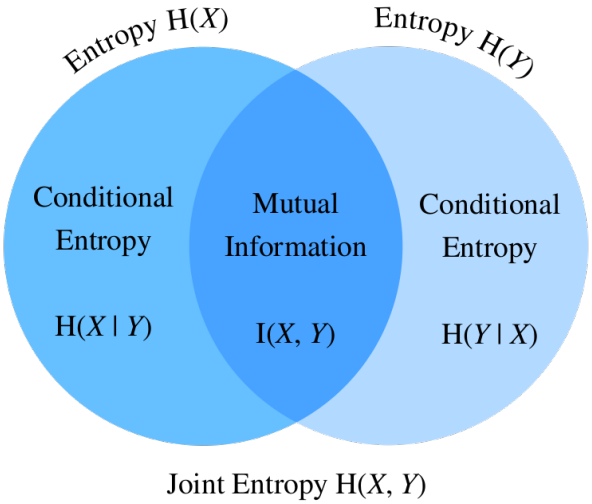


FIGURE 2.4: Joint Entropy

2.5 Network model

Network theory is the study of graphs to highlight relationships between objects. It is part of graph theory. The scope of graph theory is very broad including applications physics, statistics, computer science, biology, health, engineering, finance, sociology, etc.

The birth of graph theory is considered to be in 1736 when the Swiss mathematician Leonhard Euler solved the problem known as the seven bridges of Königsberg. The problem was finding a route in the Prussian city with which to cross all the bridges only once and then return to the starting point. Euler approached the problem in topological terms and proved that such a path did not exist.

His theorem states that any graph can be traversed in such a way that one passes only once from each edge, if and only if, all the nodes are of even degree or at most there are two of odd degree. In the latter case, it will be necessary to start from an odd node and end on the other odd node.

2.5.1 Graph Theory

Basic Concept

A graph G is a mathematical object defined as an ordered pair of sets G = (V, E), where V = v1, ..., vn is the set of vertices or nodes and E is the set of edges or arcs.

In graph G = (V, E) of size N, the minimum number of nodes is 0 and the maximum number is $N(N-1)/2$.

A graph can be simple if it does not contain self-loops (links from a node to itself) or multiple edges (nodes connected by more than one link) or not simple if it contains them.

A graph can be directed or undirected and weighted or unweighted (figure 2.5). A graph G = (V, E) is a directed graph (direct graph or digraph) if the arcs have a direction, you can go from node 1 to node 2 but not vice versa (fig.C), otherwise, it is undirected (fig.A). The graph G = (V, E, w) is a weighted graph where each edge corresponds to a weight (fig.B, D).

Newman, 2003 explains the structure and function of complex networks.

Source: (Farahani, Karwowski, and Lighthall, 2019)

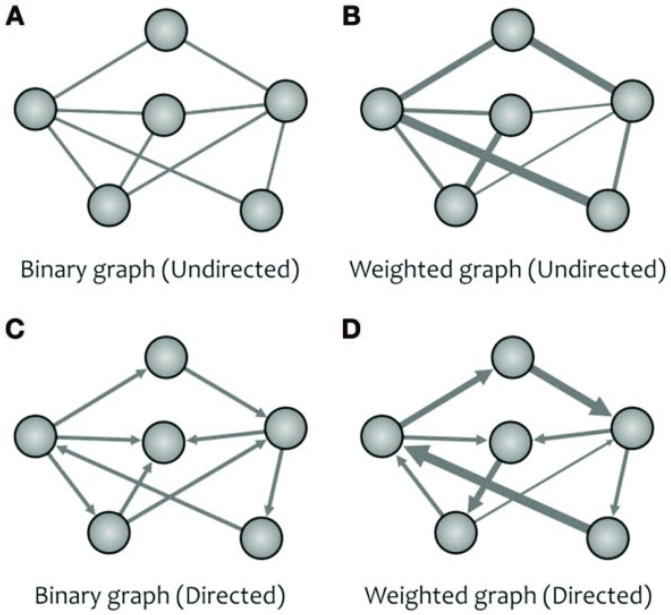
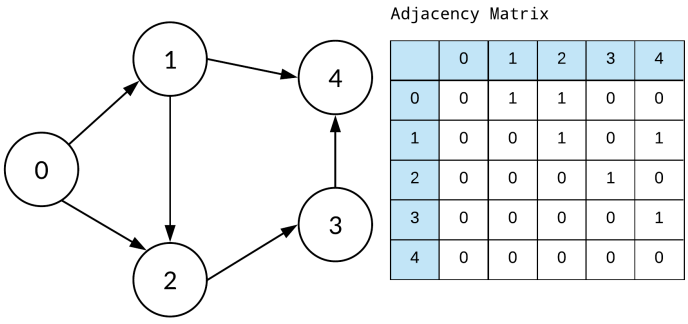


FIGURE 2.5: Graphical representation of graph

The adjacency matrix

Given any graph, its adjacency matrix is made up of a square binary matrix.
Where the rows indicate the source nodes and the columns indicate the destination nodes. Each element of the matrix indicates whether the two nodes are connected to each other (1) or not (0) (example in fig 2.6).

Source: Codepath[8]



	0	1	2	3	4
0	0	1	1	0	0
1	0	0	1	0	1
2	0	0	0	1	0
3	0	0	0	0	1
4	0	0	0	0	0

FIGURE 2.6: Adjacency matrix

Distance

The length path is the distance between two vertices u and v. The shortest path that must be traversed to go from u to v is called the geodetic path.
Two vertices are adjacent if they have a distance of 1.
The diameter of a graph is the maximum of the distances between two nodes of the graph.

Centrality measures

Centrality measures are specific measures of each node and relate, in general, to the importance of a node within the network theorized by Freeman, 1978.

The three main measures are:

- Degree centrality measures the number of connections of a node.

$$d_i = \sum_j m(i,j), \tag{2.15}$$

  where $m(i,j)$ = 1 if there us a link from node i to node j.

  For the direct graph we will talk about out-degree which represents the number of outgoing arcs and in-degree which represents the number of incoming arcs.

  The maximum degree of graph G is the maximum degree of its nodes, and the minimum degree of the graph the minimum degree of its nodes.

  A node of null degree is called an isolated node.

- Closeness centrality measures the proximity of one node to the others.

$$c_I = \sum_j d(i,j), \tag{2.16}$$

- Betweenness centrality is the measure of how much a node can connect all the other nodes calculated based on how often this is found on the shorter paths (i.e. geodetic paths) among the other pairs of nodes in the network.

$$b_i = \sum_{j,k \epsilon N,} \frac{n_{jk}(i)}{n_{jk}}, \tag{2.17}$$

where n_{jk} is the number of shortest paths connecting i and j , while $n_{jk}(i)$ is the number of shortest paths connecting i and j and passing through i.

Community Structure

Clustering is the measure of the degree to which the nodes of a graph appear to be bound to each other. A review of definitions and measures of cluster quality was written by Schaeffer, 2007. There are two measures of clustering: local and global.
Local clustering coefficient

The local clustering was defined as the cliquishness of a typical friendship circle (neighborhood) by Watts and Strogatz, 1998.

In a graph $G = (V, E)$, the neighborhood N_i of a vertex is known as its neighbors who are directly adjacent to it.

$$N_i = \left\{ v_j : e_{ij} \epsilon E, \vee e_{ij} \epsilon E \right\}, \tag{2.18}$$

where an edge e_{ij}connects vertex v_i with vertex v_j.
For directed graph:

$$C_i = \frac{|\left\{ e_{jk} : v_j, v_k \epsilon N_i, e_{jk} \epsilon E \right\}|}{k_i(k_i - 1)}, \tag{2.19}$$

where k_i is the number of neighbours of a vertex.
For undirected graph:

$$C_i = \frac{2|\left\{ e_{jk} : v_j, v_k \epsilon N_i, e_{jk} \epsilon E \right\}|}{k_i(k_i - 1)}, \tag{2.20}$$

The values of C_i ranges from 0 to 1, where 0 means that there aren't connections while 1 means that the graph is complete (clique).
Global clustering coefficient

The global clustering coefficient is also know as transitivity,

$$C = \frac{3N_{\triangle}}{N_3}, \tag{2.21}$$

where $N_{\triangle}$ is the number of closed triplets (triangles in the network) and N_3 is the number of connected triples. In the literature, a cluster is also called a community (Newman and Girvan, 2004).

Girvan and Newman, 2002 highlight a property of many networks whereby nodes are tied into joined groups. This property is the definitive property of the community structure.

The essence of identifying communities is that the nodes of the same community are very similar to each other and very different compared to the nodes of other communities. For this reason, it is very useful for large networks. In this case, the nodes of the same community have high probabilities to share features and characteristics.

Newman and Girvan, 2004 propose a measure for the quality of a division, called modularity .

2.5.2 Graphical model

So far we have outlined graphs in which we knew nodes and arcs, but in most circumstances, this is not the case. In this last event, we speak of probabilistic graphs where the nodes are known and the arcs are estimated through probabilistic rules.

A graphic model, also called probabilistic graphic model (PGM), is a probabilistic model that expresses the structure between random variables through a graph. Its main goal is to understand this structure through a graph that is a factorized representation of a set of independent variables that exist in a given distribution.

We provide here some basic definitions related to probability:

- *Joint probability* represents the probability of two (or more) variables occurring together denoted by $P(A, B)$.
- *Conditional probability* represents the probability of a variable (or more) given another variable (or more) denoted by $P(A|B)$, it assumes the knowledge of a variable that affects another.
- *Conditional independence* represents the probability that two variables are independent given another variable denoted by $P(A|B, Z) = P(A|C)$, it means that variable A is conditionally independent of variable B given an event C.

The most used models in graph theory are bayesian networks and markov networks. The main difference between the two methods is that the first is a direct model, this determines the presence of a cause-effect relationship between the variables, the second instead is indirect so this relationship does not exist.

Bayesian networks

"Bayesian networks are a combination between probability theory and graphs" (Rao and Rao, 2014).

A Bayesian network is an oriented acyclic graph, also known as directed acyclic graph (DAGs), a type of graph that does not have direct cycles. For this reason, whatever vertex we choose, one cannot go back to it by traversing the arcs of the graph.

In a Bayesian network, a graph is a representation of conditional independence about distribution. In particular, nodes represent variables and links represent conditional dependence. The structure is defined at the hierarchical level as a family tree. When a node is directly connected to a second node, the first is called *parent* and the second one is called the *child*. All nodes that can be reached through a path that starts directly from a node are said to be descendants, while, the ancestors of a node are all the nodes from which the node can be reached through a direct path. The initial and final node (*root* and *leaf* respectively) are the terminal nodes.

As the name suggests Bayesian networks are based on the Bayes theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{2.22}$$

The theorem is used to determine the probabilities of each node from conditional and prior probabilities.

A simple example with three variables are represented in figure 2.7

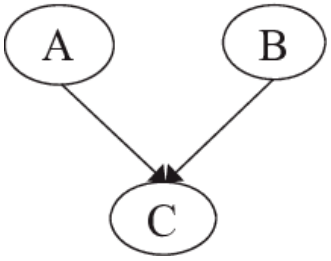


FIGURE 2.7: Simple Bayesian network with three variables

The directed edges represent a direct dependence (from A to C and from B to C), while the absence of edges represents conditional independence.

Formally:

$$P(A, B, C) = P(C|A, B)P(A)P(B), \tag{2.23}$$

In a Bayesian network, the links between nodes are expressed through a conditional probability distribution. Each child node X with n parents $(Y_1, Y_2..Y_n)$ has a conditional probability table that contains all the possible combinations of states of the child and the parents $P(X|Y_1, Y_2..Y_n)$.

Through the chain rule it is possible to calculate the joint probability distribution of the network as the product of the conditional and marginal probabilities of all nodes (Krieg, 2001):

$$P(x_1, x_2, ..., x_n) = \prod_{i=1}^{N} P(X_i|Pa_{X_i}), \tag{2.24}$$

where Pa_{X_i} indicates the set of parents of node Xi.

Markov Network

A Markov network (MN) is an undirected graph, that unlike the Bayesian networks can contain cycles. In this case, we want to capture the affinity or the mutual interactions between the two connected variables. If a system satisfies the Markov properties will talk about the Markov network.

Local Markov property states that a variable is conditionally independent of other variables given its neighbors.

Mathematically:

$$X_v \perp X_{V \setminus N[v]}|X_{N_v}, \tag{2.25}$$

where N_v is the set of neighbors of v, and $N_v = v$ is the closed neighbourhood of v.

Pairwise Markov property affirms that any two non-adjacent variables are conditionally independent given all other variables.

Mathematically:

$$X_u \perp X_v | X_{V \setminus \{u,v\}}, \tag{2.26}$$

Global Markov property affirms that any two subsets of variables are conditionally independent given a separating subset.

Mathematically:

$$X_A \perp X_B | X_S, \tag{2.27}$$

Markov network is composed of a graph G and a set of potential functions ϕ_k for each clique (Kok and Domingos, 2005). A clique, in graph, is a fully connected subgraph.

The joint distribution in Markov network is given by:

$$P(X = x) = \frac{1}{Z} \prod_k \phi_k(x_k), \tag{2.28}$$

where x_k is the state of the variables that appear in that clique and Z is a partition function defined as:

$$Z = \sum_{x \epsilon X} \prod_k \phi_k(x_k), \tag{2.29}$$

2.6 Summary

This chapter introduces the topics covered in the next chapters with their basic concepts and notions.

The topics, however, are very broad and an in-depth analysis was not possible. For this reason, I leave any further information to the references indicated. The connection of each topic to the financial sector was deliberately not mentioned in this chapter as it is elaborated in the related chapter.

Chapter 3

On the Improvement of Default Forecast through Textual Analysis

3.1 Summary

In this investigation, we apply textual analysis to augment the conventional set of account defaults drivers with new text-based variables. Through the employment of ad hoc dictionaries and distance measures, we are able to classify each account transaction into qualitative macro-categories. The aim is to classify bank account users into different client profiles and verify whether they can act as effective predictors of default through supervised classification models.

3.2 Introduction

The change in all sectors of the economy that we are witnessing in recent years is so rapid that it speaks of the fourth industrial revolution. In the era of big data, across all sectors companies' main asset has become data. There is an increasing use of data, as the use of digital technologies increases, the amount of information collected increases exponentially. As a result, firms sit upon swathes of data, but the key is being able to derive value from it. In the financial sector, data is used for multiple purposes, one of which is credit scoring. This refers to the techniques used to assign creditworthiness to a customer, thus distinguishing between "good" and "bad" clients i.e., clients who will repay their financing and those that will be insolvent. The probability that an applicant will become insolvent is determined by analysis the information available on the specific applicant (Hand and Henley, 1997). Credit scoring models are fundamental for banks to guarantee a correct forecast of default risk for financed loans, which translates into a reduction in losses and an increase in profits. There are numerous techniques for this purpose. Although nowadays most of the models in question use quantitative information, typically financial data, such information is no longer sufficient to properly profile customers in a world that is now increasingly digital. This type of information, called hard information, is contrasted by another category of so-called soft information. Soft information is the term used to indicate information obtained through textual analysis, in this case, we talk about unstructured data. Text mining arises in this context (Aggarwal and Zhai, 2012) (Gupta, Lehal, et al., 2009).

Even in today's standards, the traditional approach, which uses only hard information, is that which is widely used by firms but there is lack of studies that analyze textual information. Jiang et al., 2018 demonstrate how the use of textual data can increase the predictive power of a model, combining soft information with typically financial information analyzing the main p2p platforms in China. Groth and

Muntermann, 2011 state that exposure to intraday market risk management can be discovered through the use of text mining. Chan and Franklin, 2011 show, through the use of textual data, that the forecast accuracy of their model improves similar traditional models by 7%. The advantages and disadvantages of hard information are analyzed by Liberti and Petersen, 2019 who examine how information influences financial markets.

This work fits into the literature on textual analysis for credit scoring. The great innovation of this work is the type of textual data. Unlike most of the works in this area, we do not use social data but rather the data of bank transactions. The goal is to try to understand if buying habits can be indicators of insolvency.

3.2.1 Textual analysis for Credit Scoring Model

Cornée, 2019 demonstrates the importance of textual information for credit scoring by analyzing 389 presences of a French bank . Grunert, Norden, and Weber, 2005 analyzing German SMEs, compare a model based on financial data, one on textual data and one mixed demonstrates how the latter is the best in terms of predicting loan defaults .

Djeundje et al., 2021 use psychometric data to estimate credit scoring models and show how these data can be used to predict good or bad client. It stresses how it can be very helpful, in particular in country where financial information is not available. With the aim to improve access to credit for those people for whom it is not possible to obtain a credit history, Pedro, Proserpio, and Oliver, 2015 present a score, called Mobiscore, able to estimate the financial risk through the phone's data. From these data it is possible to extract the applicant's personality and status which have proved to be capable of predicting the default like the classical data. Iyer et al., 2016 use hard information and soft information in peer-to-peer market. They show the importance of soft information in screening borrowers, in particular for borrowers with lower credit quality. Netzer, Lemaire, and Herzenstein, 2019 use the free text that borrowers write when applying for a loan to investigate the likelihood of default. They find that, in the text, it is possible to detect intentions, emotional states and personality traits. They demonstrate how the integration of this type of data, with the financial ones, is able to improve the forecast of default up to 4.03%. In the context of peer-to-peer lending, Niu, Ren, and Li, 2019 study the impact of information obtained from social media on loan insolvency. Through three machine learning algorithms (random forest, AdaBoost, and LightGBM) they demonstrate that there is a significant correlation between such soft information and the default and that they can be used to improve default predictions. Guo et al., 2016 use a Latent User Behavior Dimension based Credit Model (LUBD-CM) for the applicant's credit analysis and show that this model significantly improves the forecast compared to standard models.

In this analysis, we propose a credit scoring model that utilizes text mining. The variables extracted through textual analysis are used as predictors in the model. To extract this information we have classified the bank transactions into macro-categories and then considered the frequencies of each macro category and the total amounts. We then compared the classical model based on financial information and the one with the addition of variables derived from textual analysis.

The rest of the analysis is organized as follows: in section 3 the methodology used is shown, in section 4 we analyze the data, in section 5 we report the results obtained and finally the paper is summarized and future research presented.

3.3 Methodology

We developed a default risk prediction model by combining financial information and textual information. The first step of the analysis was the extraction of relevant variables from the texts provided in the transactions. The method consists of 2 parts: pre-processing and knowledge extraction. The first one needed a lot of work and demanded most of the time spent on the analysis.

"Preprocessing plays a significant role in Text mining. Any task of text mining fully depends on the preprocessing step. High-quality of preprocessing always yielded superior results"(Kumar and Ravi, 2016).

In the preprocessing, we have cleaned up the texts of the transactions. We have created the textual corpus starting from the original text through the removal of stop words (typical and specific for the context), tokenization, removal of errors and stemming. We then created the document term-matrix. The matrix describes the frequency of the terms in the documents. In our case, the rows represent the transactions and the columns correspond to the terms. The columns are extracted through the analysis of the corpus regarding the dictionaries, these were created manually in terms of key-value: by inserting the subject of the service as a key and the category of the transaction as value. The final goal of text mining was to obtain topics (macro-areas of interest relating to transactions) to create new variables for each one created and to insert these variables in the credit scoring model. The first problem encountered is spelling errors, and due to this problem many transactions were not found in dictionaries. This obstacle is because the transaction descriptions are handwritten by the bank operators, who often abbreviate the words or write absently without paying attention. To overcome this issue we, therefore, decided to use a distance measure, the Levenshtein one, to attribute the word to the closest transaction under consideration (Levenshtein, 1966). This is a measure accounting for the difference between two strings, which is the minimum number of changes necessary to transform one word into another. Mathematically:

$$lev_{a,b}(i,j) = \begin{cases} max(i,j) & \text{if } min(i,j) = 0, \\ min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases} \tag{3.1}$$

The algorithm flow, accordingly, consists of:

1. Pre-processing on the single transactions and creation of a textual corpus;
2. Search in the dictionary of the key corresponding to the clean string obtained from preprocessing;
3. Assignment of the value corresponding to the key of the dictionary, that identifies the category. If this is not found, the Levenshtein distance of the string is calculated from all the keys in the dictionary and the nearest dictionary value is assigned, with a maximum threshold of 10. This means that whether the distance is greater, neither category will be assigned (at the end of the investigation, the percentage of Na is around 15%);
4. After assigning a category to each transaction of the dataset, we created the variables to be included in the credit scoring model.

The categories have been grouped into macro-categories, the macro-categories chosen for the analysis are 5:

- Non-essential goods: including expenses for goods such as shopping, travel and living.
- Essential goods: including expenses in markets, pharmacies.
- Financial services and utilities: including expenses related to banks, payment services, telco companies, petrol stations.
- Revenue: including incomes such as transfers and dividends.
- Salaries: including wages and pensions.

For each of the previous 5 variables, we have therefore created 2 further variables: frequency and the total amount spent by the client for the specific category. We have thus obtained 10 new variables through the processing of textual data.

The second step of the analysis is the application of credit scoring model. The textual-based categories created were added to the financial ones and the new dataset was used in the model. The chosen model is the lasso logistics. Lasso logistic model is a shrinkage method that allows obtaining a subset of variables that are strongly associated with the dependent variable, through regularization of the coefficients bringing them to values very close or even exactly equal to zero. Since the L1 penalty is used, the variables with a coefficient equal to zero are excluded from the model. (Hastie, Tibshirani, and Friedman, 2009). Mathematically:

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j=1}^{m} |\hat{\beta}_j| \,. \tag{3.2}$$

where y_i are the n-observations for the target variable (default/no default), x_i are the n-observations for the covariates, λ is the penalization parameter chosen by cross validation and β are the coefficient of the model.

Along with Lasso we fitted Elastic net as well. Since there were no statistical significant differences, we preferred to focus on Lasso because of an easier interpretation of the results.

We decided to use logistic Lasso for its efficiency and easy interpretability.

Before applying the lasso logistic algorithm, we pre-selected the relevant variables through the Kruskal-Wallis test. We decided to apply this test due to the size of the dataset: too few observations (400) with respect to the number of available variables (52). When the size of available data is limited, as in our case, Lasso can be not efficient enough in fitting parameters (Pereira, Basto, and Silva, 2016). Lasso is good at dropping out not significant variables if it can use an appropriate number of observations compared to the number of variables. Thus, we pre-selected the most relevant variables (without being too restrictive) through Kruskal-Wallis paying attention to the division in training and test. Kruskal-Wallis is a non-parametric method (no assumptions on the distribution of the data) that states if there is a significant difference between the groups. The null hypothesis states that the k samples come from the same population and the alternative hypothesis states that they come from different ones (Siegel, 1956).

The KW test (Conover 1971) is the non parametric version of the well known ANOVA test and represents a multivariate generalization of the Wilcoxon test for

two independent samples, that can be adapted to our problem as follows. On the basis of C independent samples (each containing the transactions of a client) of size $n_1, \ldots, n_C$ (the frequency of transactions for each client), a unique large sample L of size $N = \sum_{i=1}^{C} n_i$ is created by means of collapsing the original C samples. L can be organized as a matrix that contains a number of rows equal to N and a number of columns equal to W (the number of variables). Each entry of the matrix contains the frequency count of a specific variable along with each transaction. The KW test is then applied columns, in order to evaluate the discriminant power of each variable with respect to the client classification task. For each variable, the frequency vector corresponding to each column of the matrix L is ordered from the smallest frequency value to the largest one, and a rank is assigned to each transaction in the sample accordingly. Finally, one should calculate R_i as the mean of the ranks in each of the C original clients categories samples. The multivariate KW test can then be shown to take the following form:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{C} n_i \frac{R_i^2}{n_i} - 3(n+1) \tag{3.3}$$

After having selected only the significant variables, we applied the lasso logistics comparing the model keeping only the financial variables, the model with only the textual variables and the one obtained by the combination of the variables. For each analysis, the dataset was divided into 2 parts, 70% for training and the remaining 30% for the test. In addition, the model has been cross-validated using 5 folds. We applied the cross validation in the training set and then validated in the test set.

The comparison of the 3 models is based on mean misclassification error (mmce), area under the curve roc (auc), and roc curve (Krzanowski and Hand, 2009). Mmce is a prediction error metrics for a binary classification problem. The Roc curve is a graphical representation, along the two axes we find the sensitivity and 1-specificity, respectively represented by True Positive Rate and False Positive Rate. It is, therefore, the true positive rate as a function of the false positive rate. AUC is the area under the ROC curve, an aggregate measure of performance across all possible classification thresholds.

3.4 Data

We have undertaken this analysis starting from 2 datasets: loans and transactions relating to an Italian bank. The data are distributed as follows: 37% defaulting and 63% non-defaulting. The covering period starting from 2014 to 2018. The paper was executed in collaboration with Moneymour. Moneymour is a FinTech startup that offers a payment method to provide instant loans for online purchases. It allows client to buy immediately and pay in installments.

In the former the original variables were:

- date of the loan request,
- loan ID,
- default status,
- amount,

- number and amount of loan payments.

In the latter, there were:

- client,
- accounting date,
- value date,
- transaction amount,
- reason code and reason text.

This study analyzed 164931 transactions and 400 loans from 2015 to 2018.
The financial variables extracted are:

- sum of income,
- sum of outcome,
- average income,
- average outcome,
- number of income,
- number of outcomes,
- total number of movements,
- sum of salary and average salary

All listed variables referring to the first month, three months, six months and the previous year respectively to the request for financing, and financing obtained.

Summary statistics of financial variables are reported in Table 1.

Through the use of text mining, the transactions carried out by each client were analyzed and the new variables were created:

- salary,
- total output nonessential goods,
- total output essential goods,
- total financial services and utilities,
- total salaries,
- total income,
- frequency output non essential goods,
- frequency output essential goods,
- frequency financial services and utilities,
- frequency salaries,
- frequency income.

Summary statistics of textual variables are reported in Table 2.

The target variable is default or non default of the client defined as follows: default means the non-fulfillment of loan payment installments for 3 months in a row.

3.5 Results

In this section, we discuss the results obtained. The data set was divided into 2 portions: training and testing. 70% of the data was used for training and the remaining 30% for the test. The data in both samples were distributed as follows: 37% defaulting and 63% non-defaulting. The target variable is the default status indicated with the value 0 for the non-default and with 1 the default.

We recall that the starting dataset presented 400 observations and 53 variables. The issue regarding the high number of variables with regards to the number of observations has been overcome by selecting the most significant variables through the Kruskal-Wallis test. The variables selected after applying the test are 21 and reported in the following list:

- previous funding,
- number revenue month 1,
- number releases month 1,
- number movements month 1,
- number revenue month 3,
- number releases month 3,
- number movements month 3,
- number revenue month 6,
- number releases month 6,
- number movements month 6,
- number revenue month 12,
- number releases month 12,
- number movements month 12,
- salary,
- total output essential goods,
- total financial services and utilities,
- frequency output non essential goods,
- frequency output essential goods,
- frequency financial services and utilities,
- frequency salaries,
- frequency income.

The model chosen for the credit scoring analysis is lasso logistics which represents an efficient choice in data analysis problems like ours when a variable selection step is needed.

For greater accuracy of the metrics obtained, we conducted the analysis using k-fold cross. Moreover, we have conducted an out of sample analysis, training models on 2014-2015-2016 data and testing them on 2017 and 2018.

Parameters estimates of the 3 fitted logistic lasso models are reported in tables 3, 4 and 5 referred respectively to financial variables, textual variables and the mixed one. In particular, from table 5 we can infer that several variables both financial and textual are significant. The textual ones, of course, are of major interest being the new ones. We observe that the largest parameter is obtained by the salary flag variable: having a negative sign means that the presence of the salary on the bank account decreases the probability of default. In particular if we calculate the odds ratio we get that the probability of non defaulting is 12 times higher than the probability of defaulting. Thus, such a simple information, that can be derived by the analysis of bank transactions, can add very useful information to the credit holder. Other two textual variables are worth mentioning: frequency of income characterized by a negative sign and the total output for financial services and utilities with a positive sign. According to the former the larger the frequency of income the lower is the probability of default. Contrary, a higher number of transactions for financial services and utilities increases the probability of default. This is to say that having several incomes helps in affording financial loans but the impact of expenses for services and utilities is not negligible. Regarding purely financial variables, all of them but one shows negative signs, meaning that they reduce the probability of default. The three largest parameters are shown by 'previous funding', 'number of movements at month 1', and 'number of revenue at month 1'. What affects largely the chance of repaying loans is the presence of previous loans' request to the bank. This ensures a previous capability of respecting financial obligations. The same applies to the number of movements in the nearest month: having more movements can be considered of financial health as if we take into account the number of revenues.

Finally in table 6 we report the comparison among the 3 models: the first which considers only the financial variables, the second which considers only the textual variables and the third one which combines both. The comparison is measured through Auc and distribution of errors with particular focus on type I (False positive rate) and specificity (True negative rate) that is the opposite of false positive rate. As can be seen from Table 6, the results obtained are pretty high for the three models and even the roc curves tend to overlap many times as shown in Figure 1. On the other hand, the models are perfectly comparable in terms of AUC. Nevertheless, there are important elements, first of all, classification errors. As you can see from table 6, the mixed model and the text based one have an improvement over the type I errors. Not all errors have the same impact, some mistakes have higher implications than others. For a bank, type I error is the most dangerous one as it represents the probability of giving a loan to those who will not pay. The costs associated with type I errors are higher than type II errors, type I are actual cost while type II is the opportunity for revenues lost.

3.6 Conclusions

In this investigation, we use textual data to enhance the traditional credit scoring model. We evaluated the models performance by comparing the basic model in

which only financial variables were included, against one in which there are only those extracted through text mining and the last one containing the mix of the two types of variables. From the analysis, we conclude that the addition of textual variables is relevant in the model. Although AUC do not vary much, it should be emphasized the distribution of errors. We observe an improvement over type I error. This is a promising result that encourages to further test the methodology with a larger dataset.

We can, therefore, state that despite the small size of the dataset, the analysis carried out shows how the textual analysis can be used in a credit scoring model to improve its accuracy.

The main innovation of our work, compared to literature in this field, is the source of textual data. Usually, in this area, the text to analyze credit scoring is obtained from social media. We use a different type of textual data: bank account transactions, with the aim of investigating whether consumption habits can be significant in a credit scoring model.

Future research perspectives concern the application of the model to a larger dataset not only in terms of observations but also of variables based on textual information. More data can offer other types of information not available in the data at hand. Moreover the application of other text analysis technique like topic modeling rather than the creation of manual dictionaries to make the process more automated and therefore decrease the time spent in pre-processing, can improve even more the quality of the analysis.

TABLE 3.1: Descriptive Financial Features

	prev. funding	num. rev. month 1	num. rel. month 1	num. mov. month 1	num. rev. month 3	num. rel. month 3	num. mov. month 3	num. rev. month 6	num. rel. month 3	num. mov. month 6	num. rev. month 12	num. rel. month 12	num. mov. month 12
Min.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1st Qu.	0.00	0.00	0.75	1.00	0.00	3.00	4.00	1.00	4.00	6.00	1.00	4.00	6.00
Median	0.00	1.00	6.00	7.00	3.00	16.50	21.00	6.00	34.00	39.00	11.00	55.00	71.00
Mean	0.23	1.46	13.78	15.23	4.22	41.38	45.59	8.10	81.39	89.50	14.93	150.50	165.40
3rd Qu.	0.00	2.00	23.25	26.00	6.00	74.25	83.25	12.00	149.25	165.20	22.00	248.20	271.80
Max.	0.00	9.00	90.00	93.00	43.00	253.00	265.00	59.00	466.00	485.00	97.00	873.00	911.00

TABLE 3.2: Descriptive Textual Features

	salary	total output essential goods	total financial services and utilities	frequency output non essential goods	frequency output essential goods	frequency financial services and utilities	frequency salaries	frequency income
Min.	0.00	-31084	-224137.0	0.00	0.00	0.00	0.00	0.00
1st Qu.	0.00	-1862	-19456.8	3.75	0.00	0.00	0.00	1.00
Median	0.00	0.00	-699.5	50.00	0.00	5.50	0.00	11.00
Mean	0.31	-2511	-15107.2	107.97	47.01	94.86	7.082	21.29
3rd Qu.	1.00	0.00	0.00	152.25	39.00	142.00	6.000	26.00
Max.	1.00	0.00	0.00	647.00	717.00	944.00	118.00	237.00

TABLE 3.3: Important Variables selected by Lasso Model on Financial dataset

Variables	**Parameter Estimated**
previous funding	-0.4548
number revenue month 1	-0.1553
number releases month 1	.
number movements month 1	-0.1759
number revenue month 3	.
number releases month 3	-0.0000
number movements month 3	-0.0150
number revenue month 6	.
number releases month 6	-0.0000
number movements month 6	-0.0001
number revenue month 12	0.0028
number releases month 12	.
number movements month 12	.

TABLE 3.4: Important Variables selected by Lasso Model on Textual dataset

Variables	**Parameter Estimated**
salary flag	-2.5998
freq. salaries	.
freq. income	-0.1111
total output essential goods	0.0015
total output financial services and utilities	0.0006
freq. output non essential goods	-0.0083
freq. output essential goods	.
freq. output financial services and utilities	-0.1294

TABLE 3.5: Important Variables selected by Lasso Model on Mixed dataset

Variables	**Parameter Estimated**
previous funding	-0.4984
number revenue month 1	.
number releases month 1	.
number movements month 1	.-0.1650
number revenue month 3	.
number releases month 3	-0.0020
number movements month 3	-0.0240
number revenue month 6	.
number releases month 6	.
number movements month 6	.
number revenue month 12	.
number releases month 12	-0.0007
number movements month 12	-0.0012
salary flag	-2.5011
freq. salaries	.
freq. income	-0.0022
total output essential goods	.
total output financial services and utilities	0.0003
freq. output non essential goods	.
freq. output essential goods	.
freq. output financial services and utilities	.

TABLE 3.6: Results from Lasso Logistic regression

	Auc	False positive rate	Specificity
Financial features	0.931	0.479	0.521
Text features	0.939	0.233	0.767
Mix	0.946	0.356	0.644

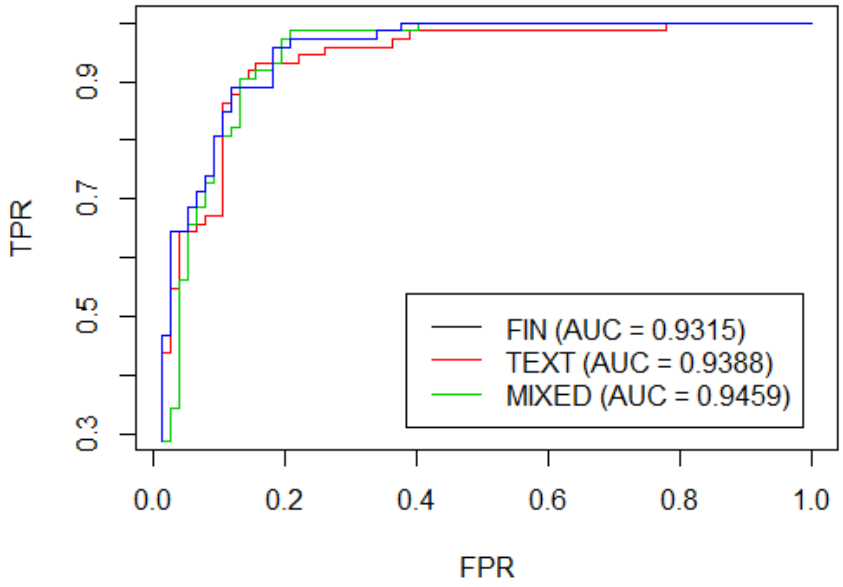


FIGURE 3.1: ROC CURVE

Chapter 4

Information Theoretic Causality Detection between Financial and Sentiment Data

4.1 Summary

The interaction between the flow of sentiment expressed on blogs and media and the dynamics of the stock market prices are analyzed through an information-theoretic measure, the transfer entropy, to quantify causality relations. We analyzed daily stock price and daily social media sentiment for the top 50 companies in the Standard & Poor (S&P) index during the period from November 2018 to November 2020. We also analyzed news mentioning these companies during the same period. We found that there is a causal flux of information that links those companies. The largest fraction of significant causal links is between prices and between sentiments, but there is also significant causal information which goes both ways from sentiment to prices and from prices to sentiment. We observe that the strongest causal signal between sentiment and prices is associated with the Tech sector.

4.2 Introduction

Causality is hard to detect from observations. This is because the occurrence of two events, one after the other, does not necessarily imply that the first caused the second. Granger, 1969 first proposed to look at causality in terms of the amount of extra information that the observation of a variable provides about another variable. In its original formulation, this corresponds to an additional term in a linear regression for financial forecasting, but the idea is general and requires the quantification of information flow between variables.

In finance, the relationships between companies are usually analyzed considering the so-called "hard" information such as stock prices, trade volumes, the quantity of output, but, in recent years, there has been an increase in the use of "soft" information including textual data, opinions, news, and sentiment. Indeed, the economic value of things and firms is both material and immaterial. Reputation is playing a major role in economics. This has probably always been true, but it has become even more crucial in the present world where social-media has a pervasive role. Therefore, a current study of market behaviour cannot be limited to the *hard* evidence related to the financial metrics but must also dig into the $soft$ metrics of social media and news. The relation between the two is still a domain in exploration.

On the one hand, an efficient market hypothesis would suggest that all information must be comprised into the prices. On the other hand, swings in social opinions

have their independent dynamics and sometimes follow, and other times anticipate market movements. In this paper, we further investigate such relationship by means of information theoretic tools, with the aim of understanding the manifest and latent dynamics of *hard* and $soft$ information within the US market.

We analyze the causality between some of the most important worldwide companies using both hard (prices) and soft (social media sentiment) information and investigate their interrelations. Causality is quantified through tools of information theory using entropy and mutual information. The first represents the uncertainty related to a variable's possible outcomes and quantifies its information content, the second one measures the information that two variables share. The transfer entropy is a conditional mutual information between the past of a variable and the future of another variable conditioned to the past of this second variable. It measures the information transferred between the two variables or equivalently the reduction in uncertainty uniquely caused by a variable on the another (Cover, 1999).

The entropy is often used in finance but usually to analyze financial information. The main innovations compared to the literature are two. The use of an information theoretic measure to monitor the causal relationship between opinions and stock market data and the choice to focus on four different networks. We analyze the causal relationships between: price to price, sentiment to sentiment, price to sentiment and sentiment to price.

4.2.1 Background: Textual Analysis in Asset Management

The use of textual analysis in the financial sector is relatively recent but constantly growing. Among the earlier papers, Engelberg, 2008 demonstrates that soft information, although more difficult to calculate, offers greater predictability on asset prices in particular at a longer horizon. Tirea and Negru, 2013 create an optimized portfolio through the combination of text mining, sentiment analysis, and risk models on the Bucharest Stock Exchange. Jothimani, Shankar, and Yadav, 2018 in their study integrate hard and soft data, the latter collected from online articles and tweets, and demonstrate that the combination of the two types of information allows optimization of the investment portfolio. Zheludev, Smith, and Aste, 2014 using sentiment techniques on social media messages show that, analyzing the S&P index, information contained in social media can impact financial market forecasts. The authors Cerchiello and Nicola, 2018 use the content of regular financial news to track the evolution across time and space of topics which are relevant in the financial context.

With a focus on the impact of negative sentiment, Tetlock, 2007, using daily content from the Wall Street journal, finds that the volume of market exchanges is determined by unusually high or low pessimistic values. Indeed, Huang, Zang, and Zheng, 2014 show that investors react differently depending on whether the information received is positive or negative; in the latter case, the reaction is stronger. They also find, on a non-market-based test, evidence that information extracted from analyst reports has predictive power on earnings growth over the following five years.

Due to the easier processing of short text data, a notable application of sentiment analysis in finance has involved the analysis of tweets. Bollen, Mao, and Zeng, 2011a examine whether the collective mood (based on six social moods: Calm, Alert, Sure, Vital, Kind, and Happy), obtained from all the tweets published in a given period in the USA, is correlated or predictive of DJIA (Dow Jones Industrial Average) values. They observe that only some of the six moods are correlated with DJIA values, with

a lag of 3–4 days. Zhang, Fuehres, and Gloor, 2011 find that, by analyzing the sentiment spikes on Twitter posts, it is possible to predict what will happen in the market the following day. Rao, Srivastava, et al., 2012 using Granger's Causality Analysis show that, in the short term, tweets influence the trend in stock prices; Ranco et al., 2015 considering 30 joint-stock companies of the DJIA index, through the "study of events" methodology (MacKinlay, 1997), a technique used in economics and finance that analyzes abnormal price changes linked to external events; for each stock, it highlights the external events grouped according to a measure of polarity. They relate the prevailing sentiment in financial tweets, in terms of volume, and stock returns showing a statistically significant dependence. Souza et al., 2015 studying retail brands analyze if there is a significant connection between sentiment and volume of tweets with volatility and return on stock prices, seeing that the data obtained from social media are relevant to understand the financial dynamics and, in particular, demonstrate how the sentiment obtained from the tweets is linked to the returns more than traditional news-wires.

You and Luo, 2013 investigate classification accuracy using textual and visual data. Carvalho, Prado, and Plastino, 2014 classify tweets through an approach where paradigm words are selected using a genetic algorithm.

Kolchyna et al., 2015 describe different techniques for classification of Twitter messages: lexicon-based method and machine learning method, and present a new method that combines the two techniques. The score obtained from the lexicon based method is the input feature for the machine learning approach, and they demonstrate that classifications are more accurate using this combined technique.

In the field of financial risk management, Cerchiello and Giudici, 2016b construct a systemic risk model with a combination of financial tweets and financial prices to comprehensively assess the impact of systemic risk.

4.2.2 Background: Information Theory

Information theory was born in 1948 with the publication of Claude Shannon's article (Shannon, 1948).

Particularly used in the financial field is the concept of entropy. Dimpfl and Peter, 2013, analyzing through entropy the flow of information between CDS (Credit default swap) and the bond market, show that information flows in both directions with the importance of the CDS market increasing over time. Kwon and Yang, 2008, using entropy, examine the flow of information between composite stock indices and individual stocks and show that this flow is stronger from indices to stocks than vice versa. Schreiber, 2000 theorizes the concept of transfer entropy as a measure of oriented coherence statistics between systems that evolve over time and Marschinski and Kantz, 2002, following this concept, analyze the flow of information between two time series: Dow Jones and DAX stock index. They introduce a modified estimator able to perform well also in the case of short temporal series. Baek et al., 2005 analyze, in the US stock market, the strength and direction of information using Transfer Entropy and conclude that companies in the energy and electricity sector influence the entire market. Nicola, Cerchiello, and Aste, 2020a analyze the US banking network, made up of the top 74 listed banks, with the aim of highlighting whether mutual information and transfer entropy are able to Granger causing financial stress indices and the USD/CHF exchange rate. For the implementation of the analysis, they used general and partial Granger causality, the latter correlated to representative measures of the general economic condition.

The main goal, in the present work, is to investigate the causal relationship between two events. We chose the asymmetric information-theoretic measure identified as transfer entropy, to detect strength and direction of transfer information between sentiment and prices. Differently from Granger Causality, we use a nonlinear estimation of the transfer entropy.

The design of the paper is organized as follows: Section 4.3 presents the methodology used, Section 4.4 presents a description of the data, in Section 4.5, we report the results, and conclusions are presented in Section 4.6.

4.3 Methods

In our work, we use a nonlinear transfer entropy estimation, first introduced in Schreiber, 2000, to identify and quantify causality between time series.

Using Shannon's measure of information (Shannon, 1948), we can denote the uncertainty associated with a variable X by:

$$H(X) = -\sum_{x} p(x) \log_2 p(x); \tag{4.1}$$

This quantity can be conditioned on a second variable to obtain conditional entropy:

$$H(X|Y) = H(X,Y) - H(Y); \tag{4.2}$$

while the information that X and Y share is instead the so-called mutual information:

$$I(X,Y) = H(Y) - H(Y|X) = H(X) - H(X|Y); \tag{4.3}$$

It expresses how the knowledge of a variable reduces the uncertainty of another, and it is symmetric in X and Y.

We can express the information transfer from X to Y in terms of conditional mutual information for a given lag k:

$$TE^{(k)}_{(X\to Y)} = I(Y_t, X_{t-k}|Y_{t-k}) = H(Y_t|Y_{t-k}) - H(Y_t|X_{t-k}, Y_{t-k}); \tag{4.4}$$

Equation (4.4) quantifies the amount of uncertainty on Y_t reduced by the knowledge of the lagged variable X_{t-k} given the information of the lagged variable Y_{t-k} itself. It is therefore a quantification of the additional information on variable Y provided by the past of variable X taking into account what is already known about the past of Y.

This expression is general and applies to either linear and nonlinear estimations. In the linear case, one uses multivariate normal modeling, in the nonlinear case, one can instead estimate Transfer Entropy with a non-parametric density estimation that directly uses the empirical frequencies of observations into histogram bins.

In this analysis, following Keskin and Aste, 2019, we adopt such a non-parametric, nonlinear approach and estimate the joint entropy using the multidimensional histogram tool available from the 'PyCausality' Python package (`https://github.com/ZacKeskin/PyCausality`). According to such method, the observation space is divided into bins and the observations are allocated to each bin depending on their value. It is evident that the appropriate choice of bins is crucial. We chose the equiprobable bins approach, which enforces that, in each bin, the number of data points is approximately the same. In previous studies Keskin and Aste, 2019, it was

shown that this approach yields the best results for artificial data where the true underlying causality structure is known. In our case, where the causality structure must be discovered, we verified that other choices, such as equi-sized bins, return similar results on our dataset; however, the equi-probable bins provide the cleanest outputs.

A limitation of this non-parametric approach is that it requires a large number of observations. Indeed, for the transfer entropy between two variables, we have to estimate a three-dimensional histogram. In general, for p variables, the dimension is at least $d = p + 1$. For any meaningful statistical analysis, the bins in the histogram must be populated and therefore one must have a number of observations that is larger than $(\text{number of bins})^d$. This method is non-parametric; however, the choice of the number of bins is important, and this could be seen as a hyper-parameter. In the present study, however, the choice is highly constrained by the sample size. We have indeed two years of observations (512 days, see Section 4.4). Therefore, the maximum number of bins should be no larger than $(512)^{1/3} = 8$. In Keskin and Aste, 2019, it was shown that results are robust for a range of different values of the number of bins. Indeed, we tested the bin number in a range between 3 and 8 obtaining consistently similar results. We eventually decided for a number of bins equal to 5, which was giving the cleanest result. It should be clear that, with this non-parametric approach, with the present dataset, it would be unfeasible to extend the analysis to greater dimensions beyond the computation of transfer entropies between two variables.

Another important choice is the lag k. We chose the first-order lag $k = 1$, since we assume that one day of delay is enough to see the effects of a variable on another. This is because, in an increasingly connected world, news spread almost immediately around the world. Similarly, the time for one event to impact another is extremely close. As robustness check, we have also tested a higher number of lags up to 5, obtaining consistent results with the one here reported for $k = 1$.

The transfer entropy returns a non-negative real value. The greater the number, the larger is the amount of information measured. However, there is no reference and the number itself, without a benchmark, is of little interest. In order to obtain such a reference, we compared it with a null-hypothesis from data sets where any causal relation is removed. Such data were obtained from the original ones by shuffling randomly the time sequence of observations. In this way, we obtained both a null-hypothesis reference and its statistics. From the mean $\langle TE_{shuffle} \rangle$ and the standard deviation $\sigma_{shuffle}$ of the shuffled transfer entropy, we computed the statistical significance of the Transfer entropy results in terms of the following Z-score:

$$Z := \frac{TE - \langle TE_{shuffle} \rangle}{\sigma_{shuffle}}. \tag{4.5}$$

The Z-score provides a distance, measured in terms of standard deviations, of the observed transfer entropy with respect to expected value for non-causally related variables. Larger Z-scores imply a value of the transfer entropy that is more significantly deviating from the values expected when the variables are not causally related, implying therefore a larger likelihood of causal relation. In this paper, we used 50 shuffles. We use the Z-score because it is a robust statistical validation that depends on minimal assumptions. We checked the quantiles as well, retrieving consistent results. However, with only 50 shuffles, the quantile measure tends to be noisier. We shuffle single entries only; therefore, we eliminate autocorrelations. Shuffling blocks instead could have produced noisier null-hypothesis transfer

entropy potentially yielding to slightly lower Z scores.

Finally, we made use of the Z-score to construct graphs of significant causal links by retaining causality links at different threshold values, namely $Z > 2$ and $Z > 3$. On the resulting networks, the community detection algorithm were applied to identify causality structures. We also compared the networks between themselves and with respect to a reference network based on news.

For a better understanding of the employed methodology, hereafter we describe the step by step analysis workflow.

Step-by-step method

1. Creation of datasets
   (a) Creation of the sentiment variables based on the Brain's indicator for each company
   (b) Acquisition of daily prices for each variable for the same period
2. Cleaning datasets
   (a) Removal of weekends and holidays from the sentiment dataset
   (b) Calculation of returns for the financial dataset
3. Transfer Entropy (TE) analysis
   (a) Calculation of the TE for each pair of variables and creation of the corresponding matrix
   (b) Calculation of the Z-score for each pair of variables and creation of the relative matrix
   (c) Selection of the pairs with a Z- score greater than 3 and relative TE
4. Network construction
   (a) Construction of the network on pairs selected according to step 3.c.

4.4 Data

In this analysis, we consider the top 50 companies of S&P. The complete list of companies with the corresponding ticker code and rank Capitalization is available in Table 5.1.

TABLE 4.1: Detailed description of the top 50 S&P with ranking.

Rank	Stock	Ticker
	Communication	
13	AT & T Inc.	T
18	Verizon Comm. Inc.	VZ
	Consumer Discretionary	
3	Amazon.com Inc.	AMZN
26	Comcast Corp.	CMCSA
14	Walt Disney Co.	DIS
19	Home Depot Inc.	HD
34	McDonald's Corp.	MCD
37	Netflix Inc.	NFLX
	Consumer Staples	
39	Costco Wholesale Corp.	COST
24	Coca-Cola Co.	KO
28	PepsiCo Inc.	PEP
10	Procter & Gamble Co.	PG
43	Philip Morris Int. Inc.	PM
30	Walmart Inc.	WMT
	Financial	
12	Bank of America Corp	BAC
5	Berkshire Hathaway Inc.	BRK.B
29	Citigroup Inc.	C
6	JPMorgan Chase & Co.	JPM
22	Wells Fargo & Co.	WFC
	Industrial	
25	Boeing Co.	BA
42	Honeywell Int. Inc.	HON
47	Union Pacific Corp.	UNP
50	Raytheon Technologies	RTX

Rank	Stock	Ticker
	Healthcare	
41	AbbVie Inc.	ABBV
31	Abbott Laboratories	ABT
36	Amgen Inc.	AMGN
38	Bristol-Myers Squibb Co.	BMY
8	Johnson & Johnson	JNJ
33	Medtronic Plc	MDT
20	Merck & Co. Inc.	MRK
23	Pfizer Inc.	PFE
46	Thermo Fisher Scientific Inc.	TMO
15	UnitedHealth Group Inc.	UNH
	Tech	
2	Apple Inc.	AAPL
44	Accenture Plc	ACN
32	Adobe Inc.	ADBE
45	Broadcom Inc.	AVGO
35	Salesforce.com inc.	CRM
27	Cisco Systems Inc.	CSCO
4	Facebook Inc.	FB
7	Alphabet Inc.	GOOGL
16	Intel Corp.	INTC
17	Mastercard Inc.	MA
1	Microsoft Corp.	MSFT
40	NVIDIA Corp.	NVDA
49	Oracle Corp.	ORCL
48	PayPal Holdings Inc.	PYPL
9	Visa Inc.	V
	Energy	
21	Chevron Corp.	CVX
11	Exxon Mobil Corp.	XOM

We analyze two different types of information: stock prices and sentiment index.

The sentiment index is provided by Brain[1]. For each day, in a period starting from November 2018 to November 2020, a sentiment value is calculated from news and blogs written in English for each and every company. A brain sentiment indicator is represented by a value ranging between -1 to 1, where -1 corresponds to a negative sentiment, 0 to a neutral sentiment, and + 1 to a positive sentiment.

The workflow of Brain Sentiment indicator is described in the box.

[1]link to the site: https://braincompany.co/

Brain Sentiment indicator workflow

1. News are collected, through APIs and news feeds, from financial media and blogs (no social media), and are assigned to a specific company by the provider.
2. The assignment is checked for correctness.
3. Calculation of sentiment score using a mixed approach:
   (a) News are classified, through syntactic rules and machine learning classifiers, into specific categories with a predefined value of sentiment.
   (b) If the previous step fails, the sentiment is calculated using a Bag of Words scheme based on a proprietary dictionary. This approach is empowered by Natural Language Processing techniques.
4. Sentiment is then aggregated at the company level.

For the same period, we have daily stock prices for each company from Yahoo Finance. Since the sentiment index is available every day, differently from market data, we exclude weekend days with regards to the former, in order to have comparable time series. This choice is conducted to make the data comparable, it not affect the analysis considering that the sentiment on Monday is, however, determined by sentiment during the weekend.

For the daily stock prices, we calculate the logarithmic return

$$L = \log(Price_t) - \log(Price_{t-1}), \tag{4.6}$$

which is a rate of change of the variable. We apply such transformation just to financial data because the sentiment index is already a stable variable in a range between -1 and 1. We performed the Anderson–Darling test and verified that all sentiment variables can be considered stationary with null-hypothesis p-values all below 5%. We perform stationary tests on log returns too, and the results are the same as for sentiment variables. We add two plots for the time series (the first for returns and the second for sentiment) and also two more images on a subset for an improved visualization (Figures: 4.8, 4.9, 4.10, 4.11). This result could be deemed as a bit surprising in the light of the COVID-19 virus outbreak started in the spring of 2020, but, as already showed in (Ahelegbey, Cerchiello, and Scaramozzino, 2021), such shock had a small impact on the overall statistics of sentiment time series.

After these pre-processing steps, we obtain a complete dataset, with values on the same scale for a total of 100 variables (50 prices log-returns and 50 sentiment index) and 515 observations (two years of work-daily data).

4.5 Results

As explained in the previous sections, we want to assess the possible causal relationship between stock price and sentiment indicator focusing on some of the largest worldwide companies. To this end, we compute the transfer entropy and the relative Z-score for all couples of variables (market price and sentiment index). We have therefore 100 variables and $100 \times 99 = 9900$ distinct couples.

The full network of causality links without imposing any restriction is too dense. The large number of links and the significant density of the graph prevent inferring useful and insightful information. A more detailed and consistent analysis is depicted in Figure 4.1, where a sub-network which retains only causal links with Z-scores larger than 3 is shown. Such a stringent score allows for the presence of the most significant links. Figure 4.1 clearly zooms in on a fraction of the connections easing the readability. In this figure, and in all others, the clockwise direction of the arcs between nodes indicates the direction of connections. Note that, despite the fact that we estimated the transfer entropy between couples of variables, from the network in Figure 4.1, we can also infer properties for the relations between a higher number of variables and assess the presence of potential confounding factors. Indeed, any larger multivariate causality structure will reveal itself as a clique in the graph and any confounding factor will form a cycle. We observe only one clique of dimension three with a directional cycle JNJ $\rightarrow$ AVGO $\rightarrow$ PM $\rightarrow$ JNJ.

For a more comprehensive understanding and readability, we report in Tables 4.2, 4.3 and 4.4 the associated Transfer Entropy values and the Z-score for each couple of stock with a Z-score larger than 3.

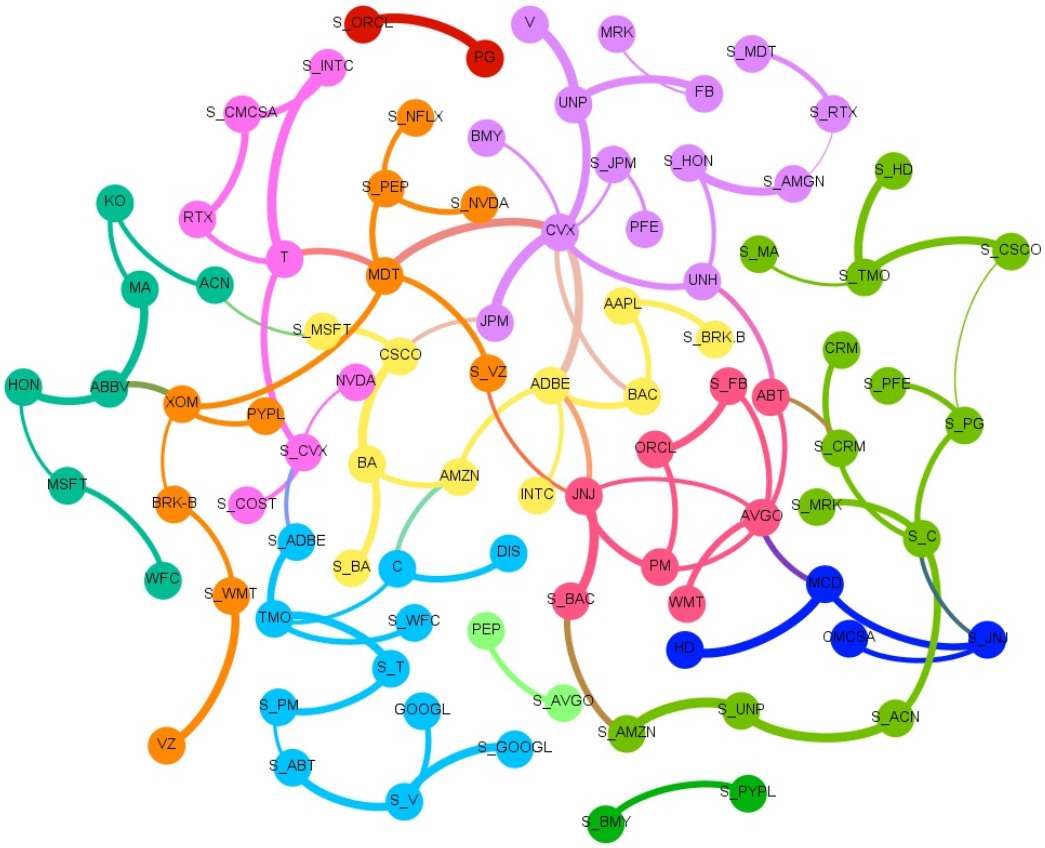


FIGURE 4.1: Network of links with Z score larger than 3. The colors represent the 12 communities found using a community detection algorithm: the Louvain method. The sentiment index timeseries is indicated with an S before the ticker's name. The clockwise direction of the curves indicates the direction of connections. Moreover, the reader can notice that there are not bidirectional interactions (to and from one vertex) and there are no cycles (paths that can be run across starting and ending with the same vertex) except for AVGO, JNJ, and PM. This result is not an imposed constraint of the algorithm but rather a result of the analysis.

TABLE 4.2: Couples of stocks with relative transfer Entropy, $TE^{(1)}_{(X\to Y)}$, values, Z scores larger than 3 (in brackets) and sectors for Price to Price network. The sectors are indicated with the capital letter, in particular we have F for Financial, H for Healthcare, T for Tech, I for Industrial, CD for Consumer discretionary, CS for Consumer staples, C for Communications, E for Energy.

Var X	**Var Y**	**Value TE (Zscore)**	**Sectors**
		Price to Price	
T	MDT	0.18 (4.24)	C→H
MSFT	WFC	0.18 (4.24)	T→F
PM	JNJ	0.18 (3.99)	CS→H
T	RTX	0.18 (3.98)	C→I
V	UNP	0.20 (3.98)	T→I
ABBV	HON	0.19 (3.81)	H→I
MCD	HD	0.19 (3.80)	CD→CD
MDT	CVX	0.19 (3.76)	H→E
UNP	FB	0.19 (3.75)	I→T
MSFT	HON	0.17 (3.66)	T→I
WMT	AVGO	0.18 (3.64)	CS→T
BAC	ADBE	0.18 (3.64)	F→T
JPM	CVX	0.20 (3.63)	F→E
UNP	CVX	0.19 (3.61)	I→E
ABBV	XOM	0.18 (3.54)	H→E
DIS	C	0.18 (3.38)	CD→F
MA	ABBV	0.19 (3.3)	T→H
C	AMZN	0.18 (3.36)	F→CD
AVGO	PM	0.1 (3.35)	T→CS
BA	CSCO	0.2 (3.35)	I→T
AAPL	BAC	0.18 (3.34)	T→F
UNH	ABT	0.18 (3.33)	H→H
CVX	ADBE	0.19 (3.33)	E→T
BRK-B	XOM	0.17 (3.26)	F→E
ORCL	PM	0.18 (3.24)	T→CS
MA	KO	0.18 (3.24)	T→CS
ADBE	INTC	0.18 (3.24)	T→T
BAC	CVX	0.18 (3.22)	F→E
ADBE	JNJ	0.18 (3.22)	T→H
C	TMO	0.18 (3.16)	F→H
FB	MRK	0.17 (3.15)	T→H
AMZN	BA	0.18 (3.13)	CD→I
MDT	XOM	0.18 (3.11)	H→E
BMY	CVX	0.17 (3.11)	H→E
PYPL	XOM	0.18 (3.10)	T→E
CSCO	JPM	0.18 (3.1)	T→F
UNH	CVX	0.19 (3.06)	H→E
ABT	AVGO	0.18 (3.05)	H→T
ACN	KO	0.18 (3.04)	T→CS
JNJ	AVGO	0.18 (3.04)	H→T
AMZN	ADBE	0.18 (3.02)	CD→T
MCD	AVGO	0.18 (3.00)	CD→T

TABLE 4.3: Couples of stocks with relative transfer Entropy, $TE^{(1)}_{(X\to Y)}$, values, Z scores larger than 3 (in brackets) and sectors for Sentiment to Sentiment. The sectors are indicated with the capital letter, in particular we have F for Financial, H for Healthcare, T for Tech, I for Industrial, CD for Consumer discretionary, CS for Consumer staples, C for communications, E for Energy.

Var X	**Var Y**	**Value TE (Zscore)**	**Sectors**
	Sentiment to Sentiment		
AMGN	HON	0.2 (4.63)	H→I
AMZN	UNP	0.2 (4.57)	CD→I
C	CRM	0.18 (4.51)	F→T
C	ACN	0.19 (4.47)	F→T
TMO	CSCO	0.19 (4.36)	H→T
AMZN	BAC	0.19 (4.34)	CD→F
BMY	PYPL	0.19 (4.3)	H→T
TMO	HD	0.2 (4.04)	H→CD
V	ABT	0.2 (3.97)	T→H
V	GOOGL	0.2 (3.89)	T→T
INTC	CMCSA	0.19 (3.82)	T→CD
ACN	UNP	0.2 (3.79)	T→I
NVDA	PEP	0.18 (3.57)	T→CS
MRK	C	0.19 (3.45)	H→F
T	PM	0.19 (3.42)	C→CS
PFE	PG	0.18 (3.33)	H→CS
ABT	PM	0.17 (3.32)	H→CS
TMO	MA	0.17 (3.32)	H→T
C	PG	0.18 (3.31)	F→CS
MDT	RTX	0.18 (3.12)	H→I
CVX	COST	0.18 (3.09)	E→CS
PEP	NFLX	0.18 (3.08)	CS→CD
JNJ	C	0.18 (3.07)	H→F
ADBE	CVX	0.18 (3.07)	T→E
RTX	AMGN	0.16 (3.04)	I→H
PG	CSCO	0.16 (3.03)	CS→T

TABLE 4.4: Couples of stocks with relative transfer Entropy, $TE^{(1)}_{(X\to Y)}$, values, Z scores larger than 3 (in brackets) and sectors for Price to Sentiment and Sentiment to Price networks. The sectors are indicated with the capital letter, in particular we have F for Financial, H for Healthcare, T for Tech, I for Industrial, CD for Consumer discretionary, CS for Consumer staples, C for communications, E for Energy.

Var X	**Var Y**	**Value TE (Zscore)**	**Sectors**
	Price to Sentiment		
JNJ	BAC	0.2 (4.37)	H→F
TMO	ADBE	0.19 (4.05)	H→T
TMO	T	0.19 (3.92)	H→C
T	INTC	0.20 (3.83)	C→ T
ABT	CRM	0.18 (3.54)	H→T
BA	BA	0.19 (3.51)	I→I
MDT	VZ	0.18 (3.36)	H→C
AAPL	BRK.B	0.18 (3.34)	T→F
BRK-B	WMT	0.18 (3.18)	F→CS
JNJ	VZ	0.17 (3.08)	H→C
GOOGL	V	0.18 (3.06)	T→T
MDT	PEP	0.18 (3.05)	H→CS
	Sentiment to Price		
CVX	T	0.19 (4.34)	E→C
ORCL	PG	0.20 (4.24)	T→CS
FB	ORCL	0.19 (4.01)	T→T
WMT	VZ	0.12 (3.83)	CS→C
WFC	TMO	0.18 (3.68)	F→H
MSFT	ACN	0.17 (3.64)	T→T
CMCSA	RTX	0.19 (3.61)	CD→I
JNJ	CMCSA	0.18 (3.41)	H→CD
AVGO	PEP	0.18 (3.38)	T→CS
JNJ	MCD	0.19 (3.37)	H→CD
JPM	PFE	0.17 (3.29)	F→H
HON	UNH	0.18 (3.19)	I→H
CVX	NVDA	0.17 (3.17)	E→T
MSFT	CSCO	0.19 (3.12)	T→T
JPM	CVX	0.17 (3.06)	F→E
CRM	CRM	0.19 (3.06)	T→T
FB	AVGO	0.18 (3.03)	T→T

The three tables report results classified according to the S&P industry sectors: Consumer discretionary, Consumer staples, Energy, Healthcare, Tech, Financial, Industrial and Communications. The sectors are not homogeneously populated, in particular, healthcare and tech ones have the largest number of stocks, respectively, 10 and 15 companies. Whilst the sector's classification is important for the correct assessment of the pattern drivers, the tendency of big companies to diversify the types of business more and more is unquestionable. As an example, Amazon, which is listed in the Consumer discretionary sector, has a division named 'Amazon Web

Services' for cloud computing and device and a division named 'Amazon Studios' for music and videos streaming. Bear in mind that the division among the sectors does not completely reflect the real connections among the companies.

A community detection algorithm (Fortunato, 2010) is employed to investigate the presence of meaningful communities inside our network in Figure 4.1.

The community detection algorithm implemented is the Louvain method (Blondel et al., 2008), a heuristic method that is based on modularity optimization. It is an unsupervised algorithm that partitions the network into mutually exclusive communities in two steps: modularity optimization with local node relocation and community aggregation. We selected this algorithm due to its simplicity and computational efficiency.

The community algorithm finds 12 different communities as we can see from the different colors. Most of the communities are similar in terms of number of companies. Interestingly, such groups have some recognizable overlap with S&P sectors, but also distinctive features revealing the different nature of market price and sentiment interconnections, which goes well beyond companies' core business.

By looking at the connections in such a network, we can distinguish between variables associated with the price returns (identified generically as 'price' hereafter) and variables associated instead with sentiment scores (identified generically as 'sentiment' hereafter).

We observe that most of the links are from Price to Price (See Table 4.2), followed by the links from Sentiment to Sentiment (see Table 4.3) and then the Sentiment to Price and finally Price to Sentiment (see Table 4.4). We observe an interesting asymmetry between companies and sectors that are influencers and the others that are followers with most of the significant links involving two different industry sectors. The leading one, in terms of number of significant links, is the Technological sector with a predominance of connection towards the Consumer sector: Accenture causing ($\rightarrow$) Coca-Cola; Mastercard $\rightarrow$ Coca-Cola; Broadcom $\rightarrow$ Philip Morris; Oracle $\rightarrow$ Philip Morris; Amazon $\rightarrow$ Adobe; McDonald's $\rightarrow$ Broadcom; Walmart $\rightarrow$ Broadcom. The influence of different sectors on the energy sector isa also interesting: Bank of America, Bristol, JPMorgan, Medtronic, UnitedHealth and Union Pacific cause Chevron; while Paypal causes Exxon. We note that this abundance of links to the energy sector is unique to this Price to Price network. There are also several links within the same sectors: a connection between United health $\rightarrow$ Abbot, both in the Healthcare sector; McDonald's $\rightarrow$ Home Depot, in the Consumer sector; and Adobe $\rightarrow$ Intel in the Tech sector.

There are also numerous links in the Sentiment to Sentiment network (see in Table 4.3). In this case, many links are related to the healthcare sector, most of them are relationships between the healthcare and the consumer sector: Johnson&Johnson $\rightarrow$ Walt Disney; Merck&Co $\rightarrow$ Walt Disney; Thermo Fisher $\rightarrow$ Home Depot; Pfizer $\rightarrow$ Procter&Gabmble; and Abbott $\rightarrow$ Philip Morris. We also find links between companies in the same sector: Pepsi $\rightarrow$ Netflix; and Walt Disney $\rightarrow$ Procter&Gamble.

In the Price to Sentiment network (Table 4.4), we notice that there is a significant frequency of stocks related to the healthcare sector which affect other sectors: tech (Thermo Fisher $\rightarrow$ Adobe, Abbott $\rightarrow$ Salesforce.com); financial (Johnson&Johnson $\rightarrow$ Bank of America); consumer (Medtronic $\rightarrow$ Pepsi); and communications (Thermo Fisher $\rightarrow$ AT&T, Johnson&johnson $\rightarrow$ Verizon and Medtronic $\rightarrow$ Verizon).

Perhaps the most interesting result lays upon the causal links from Sentiment to Price (Table 4.4). Most of them are in the technological sector in particular tech to tech: Microsoft $\rightarrow$ Accenture; Facebook $\rightarrow$ Broadcom; Salesforce.com, Microsoft $\rightarrow$ Cisco; and Facebook $\rightarrow$ Oracle.

The analysis reveals a dominant role of healthcare and technology both as influencer and follower sectors across all four networks. Another important sector is consumer, both essential (staples) and discretionary, which are, however, mainly followers and less influencers.

To ease the interpretation, we report in Figures 4.2–4.5 an aggregated network visualization of Tables 4.2, 4.3 and 4.4 representing the flows of influence between industry sectors quantified as total, significant ($Z > 3$), transfer entropy exchanged in each direction. This analysis allows for a global view of the eight sectors in terms of reciprocal influence. We note that the four networks have very distinct characteristics.

Specifically, in the Price→Price network in Figure 4.2, we observe a role of the energy sector, being a follower of both financial and healthcare sectors, a role that is not revealed in any of the other networks. Moreover, we stress that the financial sector, which traditionally plays a pivotal role when the financial market is considered, appears to be not so predominant. Indeed, the largest average Transfer Entropy is measured from healthcare to energy with 0.92. These results are in line with Ahelegbey, Cerchiello, and Scaramozzino, 2021, which showed that the healthcare sector increased the level of importance (expressed in terms of network connectivity) during the waves of the pandemic outbreak in the US market.

The Sentiment→Price network in Figure 4.3 has a major self-influencing loop with the sentiment on the technological sector affecting its own price (TE 0.92); it also reveals some influence of the financial sector on healthcare (TE 0.36) and healthcare on consumer discretionary (TE 0.37).

In the Price→Sentiment network in Figure 4.4, the main leading role is played by healthcare, and the role of the communication sector as a follower of healthcare (TE 0.55) and as an influencer of technology (TE 0.2) also emerges. This is not present in any of the other networks. healthcare is also influencing technology (TE 0.37).

Finally, the Sentiment→Sentiment network in Figure 4.5 shows a dominating role of healthcare that is affecting the consumer sectors (TE 0.56), industry (TE 0.38), and technology (TE 0.55).

Overall, the Price→Price network has the largest number of connections i.e., 25, then Sentiment→Sentiment follows with 19, finally Sentiment→Price and Price→Sentiment with, respectively, 10 and 9.

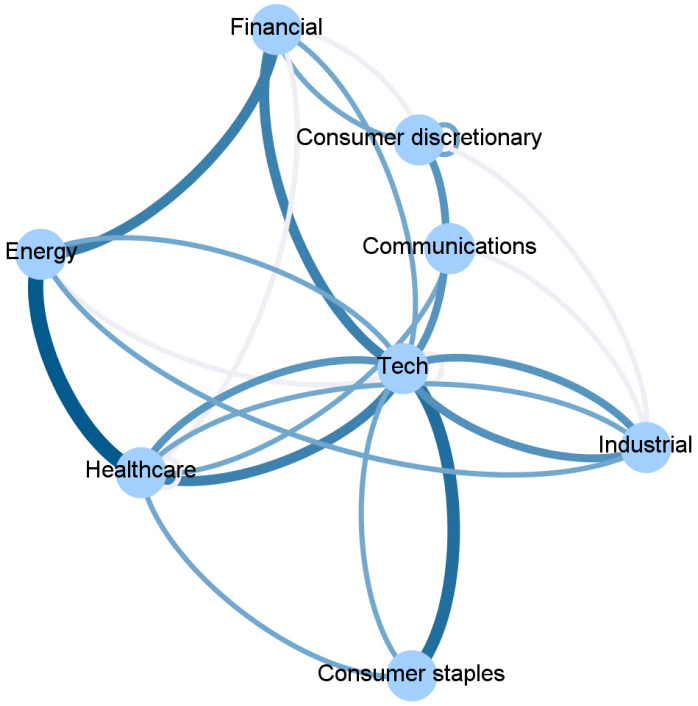


FIGURE 4.2: The aggregated Price → Price network visualization of Tables 4.2, 4.3 and 4.4 representing the flows of influence among sectors quantified as total, significant ($Z > 3$), transfer entropy exchanged in each direction. The clockwise direction of the curves indicates the direction of connections.

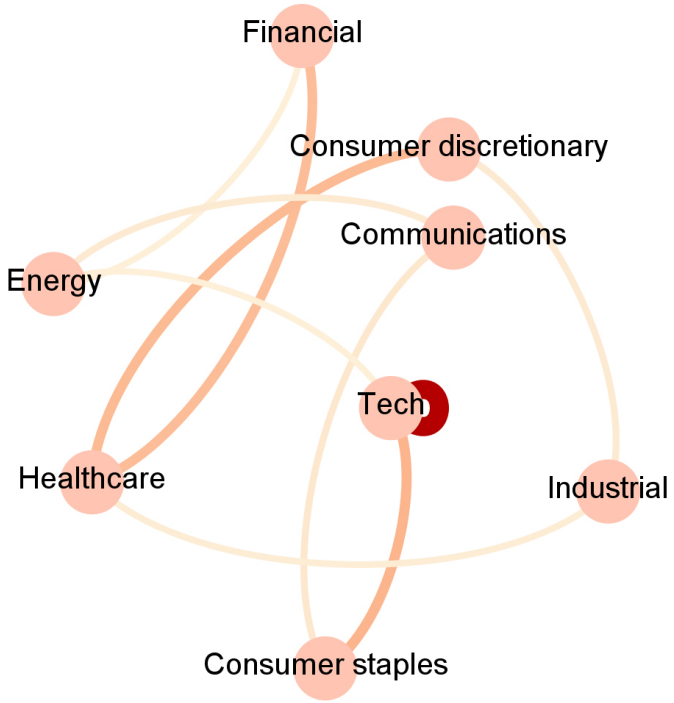


FIGURE 4.3: The aggregated Sentiment → Price network visualization of Tables 4.2 , 4.3 and 4.4 representing the flows of influence among sectors quantified as total, significant (Z > 3), transfer entropy exchanged in each direction. The clockwise direction of the curves indicates the direction of connections.

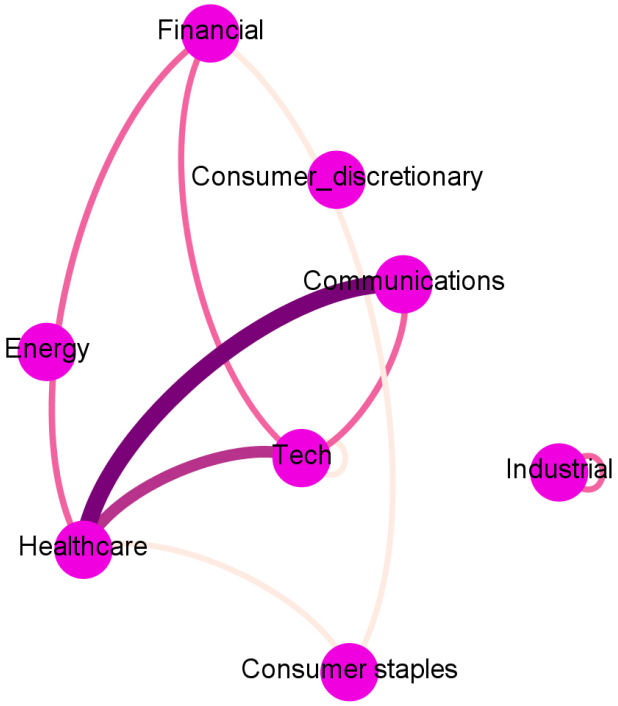


FIGURE 4.4: The aggregated Price → Sentiment network visualization of Tables 4.2 , 4.3 and 4.4 representing the flows of influence among sectors quantified as total, significant (Z > 3), transfer entropy exchanged in each direction. The clockwise direction of the curves indicates the direction of connections.

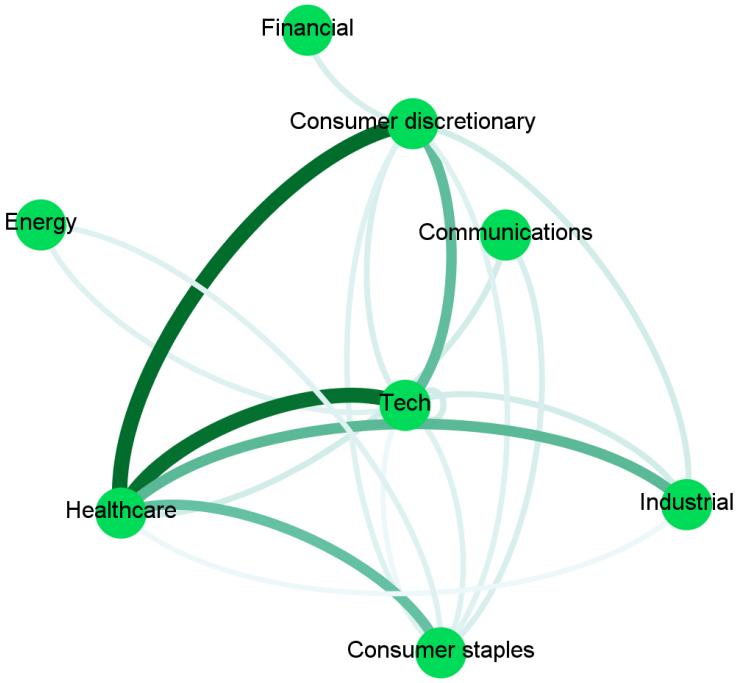


FIGURE 4.5: The aggregated Sentiment → Sentiment network visualization of Tables 4.2 , 4.3 and 4.4 representing the flows of influence among sectors quantified as total, significant (Z > 3), transfer entropy exchanged in each direction. The clockwise direction of the curves indicates the direction of connections.

4.5.1 Comparison between TE Matrix and Dataset Based on News

Since one of the main aims of our paper is to disentangle the role played by the information disclosed through news and measured by means of a sentiment score, we further analyze such component. To deepen our investigation, we pay greater attention to the sentiment aspect carrying out a further analysis using data concerning news provided by the Brain (link to the site: `https://braincompany.co/`) to identify relations between stocks by counting the number of times two tickers are mentioned within the same news article.

In Figure 4.6, we report the complete network of news in common. As already happened with unrestricted analysis, the network appears to be too dense to be readable. However, some clear patterns are already evident, like the strict connections among the company giants like AAPL, MSFT, GOOGL, FB, and AMZN (bottom right in blue), which indeed represent a community per se.

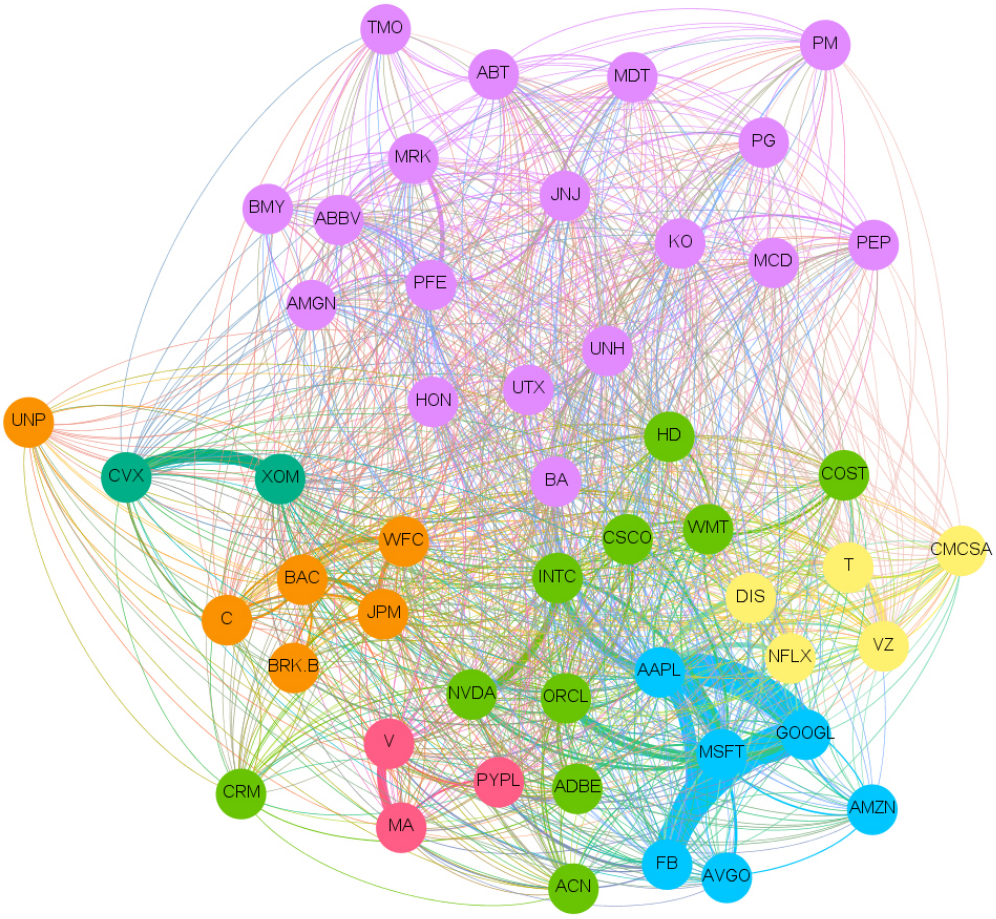


FIGURE 4.6: Network news in common. The colours represent the seven communities found using a Community detection algorithm. The clockwise direction of the curves indicates the direction of connections.

To ease the readability, we filter out the less significant links; thus, in Figure 4.7, we report the network built by retaining only the connections between stocks that score a number of news in common larger than a threshold value of 20 (such value has been identified after some sensitivity analysis).

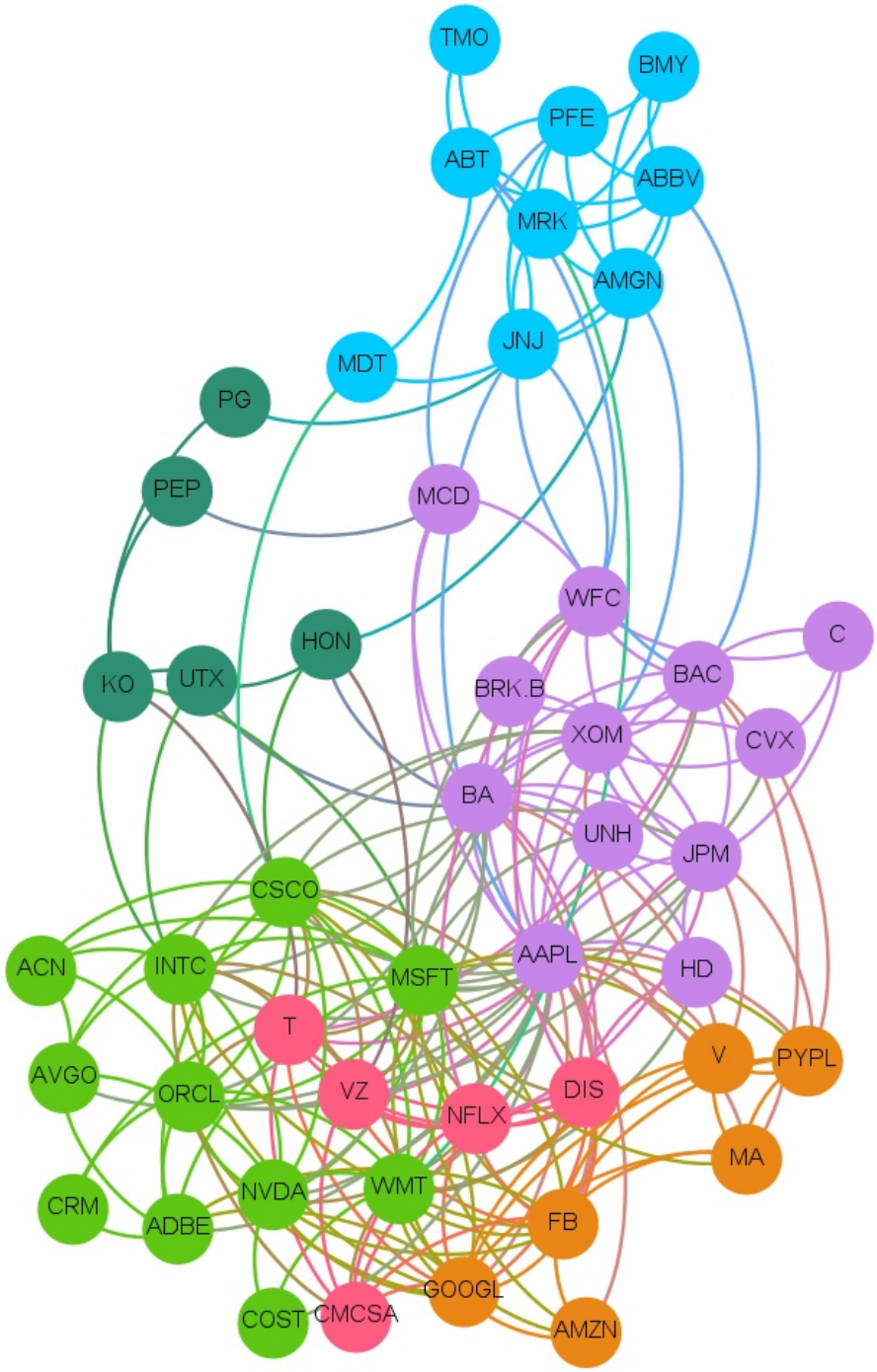


FIGURE 4.7: Network news in common larger than 20. The colours represent the seven communities found using a community detection algorithm. The clockwise direction of the curves indicates the direction of connections.

Such a network is then compared with the previous causality networks for Price to Price (PP) Figure 4.2, Sentiment to Price (SP) Figure 4.3, Price to Sentiment (PS) Figure 4.4, and Sentiment to Sentiment (SS) Figure 4.5 obtained by imposing on the links a threshold Z-score value.

Results for the thresholds: $Z > 2.5$ and a number of news in common larger than 20 are reported in Table 4.5. The reader can see that there is a rather modest overlap

between the networks that mostly involve very popular companies.

In order to statistically quantify the significance of such overlap between the networks, we compute the hypergeometric probability to have a certain number or more of overlapping edges in two directed graphs. Of course, results depend upon the chosen thresholding for the number of news and the Z-score. Overall, we find that there is no statistical significance in terms of p-value for the thresholds $Z > 2.5$ and News > 20. However, this does not mean that the links are just by chance.

By performing a sensitivity analysis by changing the threshold values, we observe that the four networks have different patterns. The Price to Price causality network shows relations with news with a rather large number of overlaps and statistical significance with p-values below 1% but only when the network is less restricted using a small news threshold and small Z-scores. This seems to indicate that news pick some insights of the internal dynamics of the market and that identify correctly important events in the financial domain that trigger propagation of information through the social media. This significance at small thresholds could indicate that this happens on average, but the importance of the news or the intensity of the causality relation is not relevant.

For what concerns the other networks, we observe that larger thresholds (more restrictive condition and less links) for the number of news in common increase statistical significance. This could indicate that news are identifying events that also resonate on the social media, but this tend to happen only for events with high relevance.

TABLE 4.5: Overlap between links in news network and links in Transfer Entropy matrix with a threshold on news equal to 20 and on Z-score equal to 2.5.

var_x	**var_y**
Price to Price variables (PP)	
NVDA	BA
BAC	AAPL
CMCSA	T
CSCO	BA
CSCO	NVDA
CSCO	ORCL
HD	JPM
INTC	T
JPM	CSCO
BA	NVDA
NVDA	MSFT
PYPL	JPM
PYPL	MSFT
WFC	MSFT
Price to Sentiment variables (PS)	
JNJ	BAC
ABBV	BMY
AAPL	BRK.B
BAC	BRKB.B
T	INTC
GOOGL	V
CSCO	WMT
Sentiment to Price variables (SP)	
MSFT	AAPL
MSFT	ACN
MSFT	CSCO
MSFT	GOOGL
CRM	ORCL
MSFT	PYPL
ABT	TMO
Sentiment to Sentiment variables (SS)	
FB	ADBE
INTC	CMCSA
ADBE	CRM
INTC	CSCO
V	GOOGL
AMGN	HON
FB	V
XOM	MSFT

4.6 Discussion and Conclusions

In this work, we study the causal relationships between opinions reflected on blogs and media and the patterns in stock market values, in order to investigate causal interactions between these variables. We focus on top 50 companies of the S&P index rooted in different sectors: consumer discretionary, consumer staples, energy, healthcare, tech, and financial industrial and communications. Data cover two years from November 2018 through November 2020. In our analysis, we employ an information-theoretic measure, the transfer entropy, to monitor the information flows between sentiment and market movements. We use a recently developed nonlinear methodology (Keskin and Aste, 2019) that can better capture causality extending the traditional Granger approach.

The main contributions of our work to literature are twofold. First, to investigate relationship between financial information and textual information through entropy using a recent nonlinear methodology and, also, to focus on different sectors: price to price, sentiment to sentiment, price to sentiment and sentiment to price.

Our information-theoretic analysis revealed a large number of strong connections. As expected, the highest number of significant causal relationships between companies involves the same kind of data source (price $\rightarrow$ price, sentiment $\rightarrow$ sentiment), but there are also strong connections across different data sources.

Some sectors are more influential in terms of sentiment dynamics and less in terms of price dynamics. For instance, in the sentiment to sentiment network, we can clearly spot the pivotal role of the healthcare sector which influences both the consumer discretionary and the technological sectors. Such pattern is present, although with differentiated importance within the other networks too. What is surprising is the role of the financial sector, which is traditionally in a paramount position compared to other sectors. Our analysis shows that financial companies are still important if we restrict to price data solely or if we consider the impact of sentiment on price but much less within the alternative scenarios. However, this is in line with what was already reported in Aste, Shaw, and Di Matteo, 2010 where a reduction of centrality of the financial sector was pointed out. This was also reported by Ahelegbey, Cerchiello, and Scaramozzino, 2021, where, through a temporal dynamic network analysis, the authors show that the financial sector behaves differently as an isolated cluster which reacts mainly to market price data (more on such peculiar pattern in Cerchiello and Giudici, 2016a). Another important sector is the technological one, either as influencer or follower depending on the network we may consider. The remaining sectors seem less consistent and change in relevance and role across the different networks.

From this study, we can conclude, first of all, that mutual influences between various companies are not limited to influences between companies within the same sector. On the contrary, the cross sector interactions tend to be more relevant. This might be because companies with high capitalization tend to operate in many markets other than their core business. Secondly, the price variables show a more homogeneous behavior, with connections which tend to be stronger and also more frequent. Nonetheless, we identify several cases where sentiment about a company has a strong influence on sentiment on other companies and also to other company prices. In particular, the tech sector reveals a very strong influence of sentiment on prices. This might be a consequence of the presence of the most popular companies in terms of branding, the 'Big Five' (Google, Amazon, Facebook, Microsoft and Apple), which are often mentioned in news and blogs and this continuous notoriety obviously affects the financial aspect.

Considering our results we can stress some possible actions for industry. The tech sector proved to be the one most influenced by sentiment, considering that, the tech companies could invest more in publicity and in initiatives aimed at improving image and therefore public opinion. The energy sector appears to be influenced by different sectors but in Price to Price network, suggesting to pay attention to price trends of the other companies. In the Sentiment to Sentiment network, the financial sector appears the most isolated suggesting to work more on communications.

The present paper can be improved and extended into several directions: US companies should be complemented and compared with European ones which typically show different patterns and level of connectedness.

TABLE 4.6: Aggregated network for the following influencing sectors: Tech, Communications, Consumer Discretionary and Consumer Staples.

Source	Target	P→P	S→S	S→P	P →S
Tech	Consumer staples	0.72	0.18	0.39	0
Tech	Healthcare	0.54	0	0	0
Tech	Financial	0.54	0	0	0.19
Tech	Industrial	0.37	0.20	0	0
Tech	Energy	0.18	0.18	0	0
Tech	Tech	0.18	0.20	0.92	0.18
Tech	Consumer discretionary	0	0.19	0	0
Tech	Communications	0	0	0	0
Communications	Healthcare	0.19	0.20	0	0
Communications	Industrial	0.18	0	0	0
Communications	Tech	0	0	0	0.20
Communications	Consumer staples	0	0.19	0	0
Communications	Communications	0	0	0	0
Communications	Consumer discretionary	0	0	0	0
Communications	Financial	0	0	0	0
Communications	Energy	0	0	0	0
Consumer discretionary	Tech	0.37	0.37	0	0
Consumer discretionary	Consumer discretionary	0.20	0	0	0
Consumer discretionary	Financial	0.19	0.19	0	0
Consumer discretionary	Industrial	0.18	0.20	0.19	0
Consumer discretionary	Consumer staples	0	0.18	0	0
Consumer discretionary	Communications	0	0	0	0
Consumer discretionary	Healthcare	0	0	0	0
Consumer discretionary	Energy	0	0	0	0
Consumer staples	Healthcare	0.19	0	0	0
Consumer staples	Tech	0.19	0.16	0	0
Consumer staples	Communications	0	0	0.20	0
Consumer staples	Consumer discretionary	0	0.18	0	0
Consumer staples	Consumer staples	0	0	0	0
Consumer staples	Financial	0	0	0	0
Consumer staples	Industrial	0	0	0	0
Consumer staples	Energy	0	0	0	0

TABLE 4.7: Aggregated network for the following influencing sectors: Financial, Healthcare, Industrial and Energy.

Source	Target	P→P	S→S	S→P	P →S
Financial	Energy	0.56	0	0.17	0
Financial	Tech	0.19	0	0	0
Financial	Consumer discretionary	0.18	0	0	0
Financial	Healthcare	0.18	0	0.36	0
Financial	Consumer staples	0	0	0	0.18
Financial	Communications	0	0	0	0
Financial	Financial	0	0	0	0
Financial	Industrial	0	0	0	0
Healthcare	Energy	0.92	0	0	0
Healthcare	Tech	0.36	0.55	0	0.37
Healthcare	Industrial	0.19	0.38	0	0
Healthcare	Healthcare	0.18	0	0	0
Healthcare	Consumer discretionary	0	0.56	0.37	0
Healthcare	Consumer staples	0	0.36	0	0.18
Healthcare	Communications	0	0	0	0.55
Healthcare	Financial	0	0	0	0.20
Industrial	Tech	0.39	0	0	0
Industrial	Energy	0.20	0	0	0
Industrial	Industrial	0	0	0	0.19
Industrial	Healthcare	0	0.16	0.18	0
Industrial	Communications	0	0	0	0
Industrial	Consumer discretionary	0	0	0	0
Industrial	Consumer staples	0	0	0	0
Industrial	Financial	0	0	0	0
Energy	Tech	0.19	0	0.17	0
Energy	Communications	0	0	0.19	0
Energy	Consumer staples	0	0.18	0	0
Energy	Consumer discretionary	0	0	0	0
Energy	Financial	0	0	0	0
Energy	Healthcare	0	0	0	0
Energy	Industrial	0	0	0	0
Energy	Energy	0	0	0	0

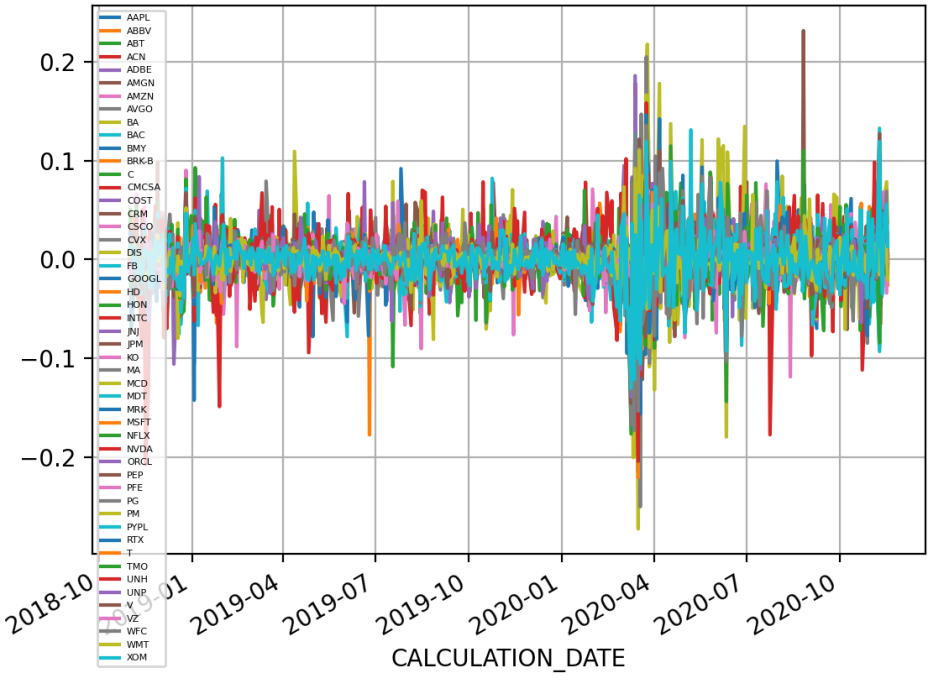


FIGURE 4.8: Time series plot of price variables (returns).

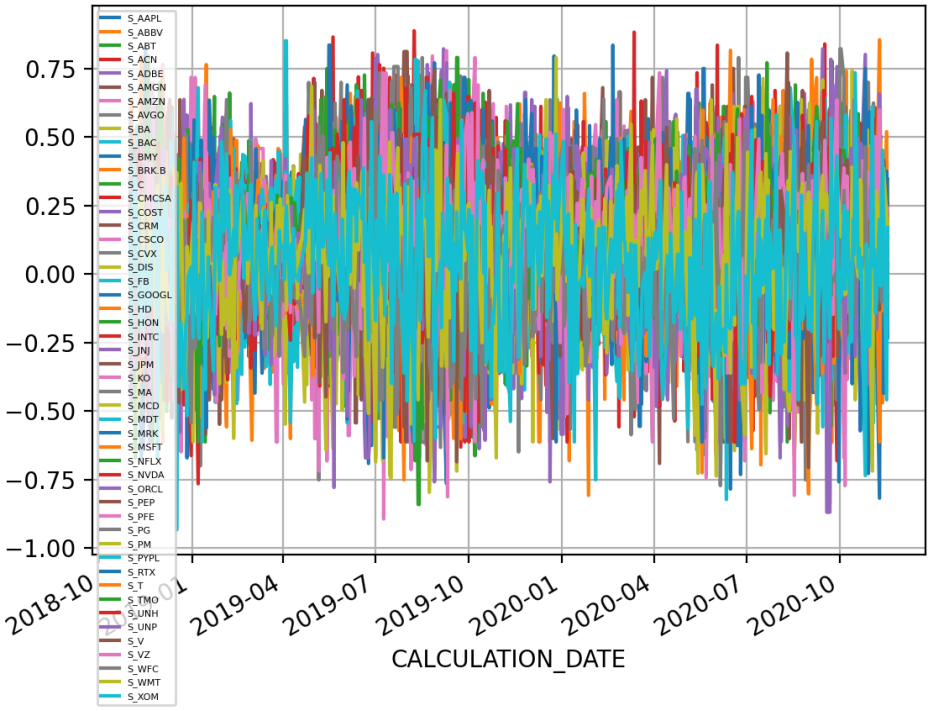


FIGURE 4.9: Time series plot of sentiment variables.

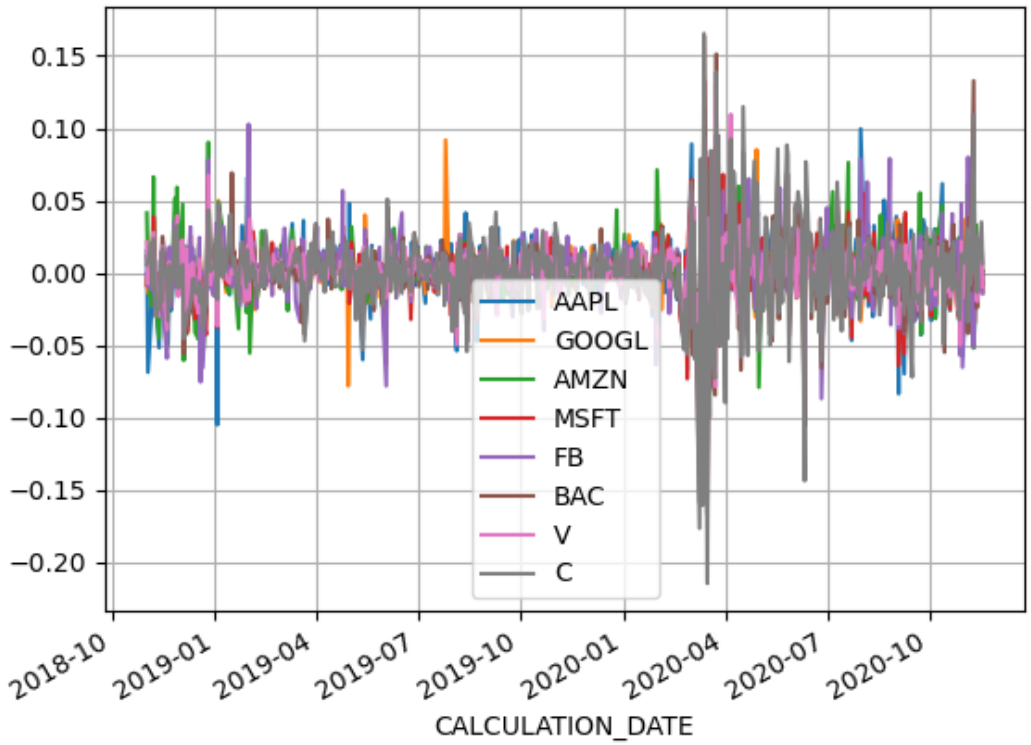


FIGURE 4.10: Time series plot of subset of price variables.

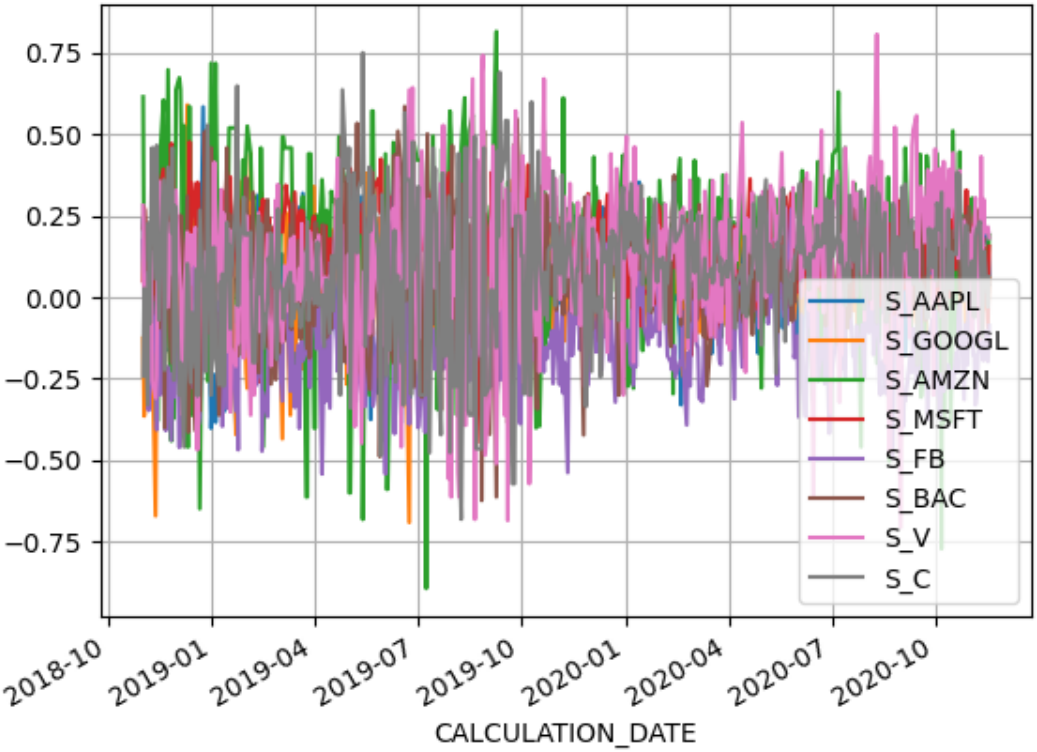


FIGURE 4.11: Time series plot of subset of sentiment variables.

Chapter 5

Network Based Evidence of the Financial Impact of Covid-19

5.1 Summary

How much the largest worldwide companies, belonging to different sectors of the economy, are suffering from the pandemic? Are economic relations among them changing? In this analysis, we address such issues by analysing the top 50 S&P companies by means of market and textual data. Our work proposes a network analysis model that combines such two types of information to highlight the connections among companies with the purpose of investigating the relationships before and during the pandemic crisis. In doing so, we leverage a large amount of textual data through the employment of a sentiment score which is coupled with standard market data. Our results show that the COVID-19 pandemic has largely affected the US productive system, however differently sector by sector and with more impact during the second wave compared to the first.

5.2 Introduction

Covid-19 is not the first pandemic that the world has experienced but the conditions we are in, changed our lives permanently and have consequences in every field. If, from the social point of view, the generated change is visible, the impact triggered at the macroeconomic level needs some time to be appreciated and quantified. As never before, the "on-line" life is so intensive, the entire world has the urgency to communicate. From the perspective of the economic market, as information spreads out, the associated sentiments change and increase the impact on the market trends. In the era of social networks, the information moves instantaneously and can amplify or damper the dynamics of the financial markets. Not only purely financial information that has an impact on the economic trend, but the price is also more and more affected by the sentiment of people. The mechanisms involved are many, from the purely economic aspects to the sociological and psychological ones. This is the reason why, at the beginning of the 2000s, sentiment analysis has been developed and largely employed, involving different sectors from marketing to politics, passing through psychology and finance.

This paper focuses specifically on the latter and, indeed, it is well known that market prices originate from complex interaction mechanisms that often reflect speculative behaviours, rather than the fundamentals of the companies to which they refer. Market models and, specifically, financial network models based on market data may, therefore, reflect spurious components that could bias results and relative discussion. This weakness of the market suggests to enrich financial market data

with data coming from other, complementary, sources. It is a fact that, market prices represent only one source of information, used for evaluating financial institutions; other relevant ones include ratings issued by rating agencies, reports of qualified financial analysts and opinions of influential and specialized media. Most of the previous sources are private, not available for data analysis. However, summary reports from them are now typically reported, almost in real time, in social networks and, in particular, in Twitter and Stocktwits.

Hereafter, we aim at investigating how and how much the interconnections among largest USA companies (top 50), have been impacted, modified and eventually reshaped, both from the financial market and the public sentiment perspectives because of the COVID-19 pandemic outbreak. To achieve the full and deep understanding of the market reactions to external shocks, we take advantage of advanced graphical models to efficiently estimate the interconnections among companies leveraging and comparing the two data sources. We completely exploit the temporal dimension by using appropriate rolling windows that reflect the market dynamics and the public perception shaping mechanism. Moreover we compare the pre-Covid-19 pandemic period with the still ongoing one, considering the 2 waves of the outbreak which have affected the USA. Data are updated till the second phase of the pandemic, namely November 17th, 2020. Our results clearly show a number of interesting facts:

- Financial market data and sentiment based data induce different behaviours in the networks structure either before and during the pandemic;
- The density of the networks evidently increases with the outbreak of the pandemic suggesting that an exogenous and rather homogeneous diffused shock produces more interconnections among the entities which may lead to a more vulnerable financial system in terms of systemic risk;
- It appears the system shows a certain amount of resilience as the first wave comes but, with the second one, the interconnections among the agents change significantly;
- A difference is evident among the sectors that reacts in their own ways, considering the relative core business and the role played in the pandemic.
- The shock produces an effect in the positioning of the companies within the network: hub and authority scores experience not only a change in the top 5 rankings but also the appearance of now comers.

This paper contributes to recent literature that analyze the economic impact of Covid-19 pandemic through text analysis. The main innovation of our research is the use of an advanced network model able to leverage the temporal-dynamic dimension of the phenomenon, considering, not only financial information but also sentiment data extracted from news.

The work is organized as follows: Section 5.3 presents the literature review, Section 5.4 presents the network VAR model and discusses the Bayesian estimation mechanism. Section 5.5 presents a description of the data, in Section 5.6 we report the results and in Section 5.7 we discuss our findings.

5.3 Literature Review

Numerous studies analyze the impact of sentiment in finance. Important papers on the statistical/econometric analysis of non conventional data are available: see,

for example, (Bollen, Mao, and Zeng, 2011b), (Bordino et al., 2012), (Choi and Varian, 2012), (Feldman, 2013b), (Cerchiello and Giudici, 2016d), who all show the added value of tweets and, more generally, of textual data, in economics and finance. Loughran and McDonald, 2016 review textual analysis literature in accounting and finance, Tetlock, Saar-Tsechansky, and Macskassy, 2008 find that language content is able to capture relevant information, not otherwise captured, which is incorporated into stock prices quickly. Cerchiello and Giudici, 2016b demonstrate how tweet data can be relevant in determining systemic risk networks and stress that such type of data has the great advantage of being able to include even unlisted institutions in the networks.

Aste, 2019, analyzing the cryptocurrency market, demonstrates how prices affect sentiment and vice versa, with differences in intensity and number of significant interactions. Souza et al., 2015, analyzing listed retail brands, demonstrate through twitter analysis that social media are very important in financial dynamics even in comparison to more traditional news sources such as newspapers. Tetlock, 2007 analyzes the link between media and stock market pointing out pessimism and demonstrating the relationship between pessimism and a decrease in stock prices and pessimism and an increase in trading volume. Joshi, N, and Rao, 2016 study the relationship between news and stock trends noting that the polarity of news (positive and negative) impacts the market. Ranco et al., 2015 analyze relationships between 30 stock companies from Dow Jones Industrial Average (DJIA) index and the blogging platform Twitter and find a significant dependence particularly during the peaks of Twitter volume.

Algaba et al., 2020 recently presented an overview of sentiment analysis related to the econometric field calling this specific research stream "sentometrics". Larsen and Thorsrud, 2019, using textual analysis on a Norwegian newspaper, construct a new index and prove that it can be useful to predict key quarterly economic variables, including assets.

Our paper supports the recent literature on the impact of Covid-19 using text analysis. We focus on the most recent papers which takes explicitly into account the effects of Covid-19 pandemic. Costola et al., 2020 examine the relationship between stock market reactions and news of COVID-19 obtained from three platforms: MarketWatch.com, Reuters.com, and NYtimes.com. They report a positive association between sentiment score and market returns and illustrate this result also applying principal component analysis on the sentiment database showing that the first principal component is positively related to the financial market. Looking at the Bitcoin market, Chen, Liu, and Zhao, 2020 study the impact of fear sentiment, affected by pandemic, on Bitcoin prices in a period from 15 January 2020 to 24 April 2020, using vector autoregressive (VAR) models and show that the fear related to pandemic channels to negative Bitcoin returns and high trading volume. Using the twitter platform, Derouiche and Frunza, 2020 study the relationship between tweets sentiment, related to sports companies and their stock prices using the Granger causality test of tweets on stocks and the event study related to Covid-period. Valle-Cruz et al., 2020 analyze the link between some twitter accounts and financial indices. They show that the market reaction is delayed by 6–13 days after the information publication and that the link between these two actors is very high. Conducting a statistical analysis of 13 million tweets for 2 weeks, Yin, Yang, and Li, 2020 note a stronger ratio of positive sentiment than negative one with particular attention to some specific topics such as "staying at home". Rajput, Grover, and Rathi, 2020, considering tweets from January 2020 until March 2020, show that most of the tweet are positive, only about 15% negative.

Considering the Italian stock market, Colladon et al., 2020 propose a new textual index (ERKs) able to predict stock market prices and demonstrate the improvement using a forecasting model. Mamaysky, 2020 examining the financial markets, note that until mid-March 2020 the markets are hypersensitive, that is very volatile and overreacting to news. From mid-March on-wards, the markets show a structural break reducing largely the hypersensitive trait. Gormsen and Koijen, 2020, analyzing equity market and dividend futures, show how these move in response to investors' expectations of economic growth. They note that the programs implemented by governments have not improved growth expectations in the short term. Baker et al., 2020, analyzing the previous pandemics (1918, 1957 and 1968), show how the Covid-19 pandemic has unprecedented effects on the US market. The authors note that this is imputable to government restrictions on commercial activities and social distancing. The socioeconomic effects of Covid-19 on every aspect of the economy have been reviewed by Nicola et al., 2020 and Zhang, Hu, and Ji, 2020 map general risk patterns and systemic risks in markets around the world. We pay particular attention to the recent literature that has studied the impact of the pandemic on the US market with a specific focus on sub sectors specificity. Lee, 2020 explores the correlation between sentiment score and 11 sector indices of the US Market through a set of t-test with different lags. Results demonstrate that all sectors present a significant boost in volatility due to the pandemic. All sectors experiences increased volatility due to the pandemic, in particular, consumer, industrial, energy and communication services are in the group of the high-medium level of correlation, utility sector in the low-level group, while tech and healthcare in the high, medium, and low group. The impact of Covid-19 was stressed also by the U.S. Federal reserve in some notes [1]. Chen, Liu, and Zhao, 2020 show the "disconnection" between stock market and real economy. High price stocks, in particular tech stocks (Facebook, Amazon, Apple, Netflix, and Google), have performed better throughout the pandemic while low price stocks performed worse, losing the 10% of their values pre-pandemic. Ahmed et al., 2020 analyze the impact of Covid-19 on Emerging Market Economies (EMEs) in particular relationship between pandemic outcomes and financial developments considering 22 financial indicators. They show that the access of EMEs to international capital markets is determined by the spread of the virus and in particular by the lockdown measures adopted to deal with it, rather than by the strength of their economies.

5.3.1 Background: Network models

We studied the impact of Covid-19 on stocks' relationship through the application of a network model. Boccaletti et al., 2006 review the structure of the networks and the applications in the different fields. Related to the financial area, Pantaleo et al., 2011 build a network structure based on covariance estimators to improve the portfolio optimization. Peralta and Zareei, 2016 propose a portfolio optimization strategy through network-based method in which the securities are the nodes of the network and the links are the correlations of returns. Pozzi et al., 2008 compare the stability of two graph methods: the Minimum Spanning Tree and the Planar Maximally Filtered Graph using financial data.

Network models approach are commonly used in the field of systemic risk. (Sheldon, Maurer, et al., 1998), (Upper and Worms, 2004), (Eisenberg and Noe, 2001) and in particular, frequently, are based on correlations between agents. There is a myriad

[1] https://www.federalreserve.gov/econres/notes/feds-notes/the-effects-of-covid-19-as-reported-by-local
htm

of studies on the application of network models to uncover these vulnerabilities in financial systems to identify channels of shock transmission among financial institutions and markets (Battiston et al., 2012)(Elliott, Golub, and Jackson, 2014)(Acemoglu, Ozdaglar, and Tahbaz-Salehi, 2015b)(Billio et al., 2012)(Diebold and Yilmaz, 2014)(Cerchiello, Giudici, and Nicola, 2017)(Nicola, Cerchiello, and Aste, 2020b)

Mantegna, 1999, studying daily time series, finds a hierarchical arrangement between them through the construction of a graph calculated on the correlations between each pair of actions. Onnela, Kaski, and Kertész, 2004 construct a network using return correlations and explain the methodology for constructing asset graphs. Giudici and Abu-Hashish, 2019 propose a correlation network VAR model to explain the structure between bitcoin prices and classic asset. Steinbacher, Steinbacher, and Steinbacher, 2013 study network-based model of credit contagion related to the banking system to analyze the effect of shocks to the financial system. Billio et al., 2012 construct a Granger-causality networks on hedge funds, banks, broker/dealers, and insurance companies showing that banks are the most important actor in transmitting shocks than others, Giudici and Spelta, 2016 improve financial network model applying Bayesian graphical models and dynamic Bayesian graphical models.

Ahelegbey, Billio, and Casarin, 2016b propose a Bayesian graphical VAR (BGVAR) model to identify channels of financial interconnectedness for systemic risk analysis. Bouri et al., 2018 apply the BGVAR model to examine the predictive power of implied volatility in the commodity and major developed stock markets. Souza and Aste, 2019 demonstrate the predictability of future stock market using a network approach that combines textual information and financial data. Giudici, Hadji-Misheva, and Spelta, 2020 propose a model for improving the credit risk of peer-to-peer platforms by exploiting the topological information embedded in similarity networks.

This work contributes to the literature specifically devoted to the assessment of the impact of the Covid-19 pandemic by means of network models. In particular, Shahzad et al., 2021b recently analyzed the reactions of the US stock market. They extend the Diebold-Yilmaz spillover index to quantile analysis showing that the network of sectorial connections is solid but changes in reaction to unfavorable market shocks but also to favorable ones. In the case of the pandemic period, the structure of the network has undergone a change leading to a closer connection between the closest clusters. In a further work, Shahzad, Hoang, and Bouri, 2021 still focuses on the US market with a specific interest on the tourism sector. They analyze 95 US tourism companies through tail risk spillover analysis showing that the whole sector was negative affected by the crisis. However, they found an increase in the importance of the small tourism firms. In both cases, the analysis of the pandemic impact is achieved through network models. The main difference with our work lies in the type of data as we consider explicitly public sentiment measured through textual data. Given the huge impact of the pandemic on people daily life and the high levels of uncertainty about the future and the evolution of the pandemic itself, we deem crucial to take directly into account opinions and sentiments.

The impact of the crisis on the Chinese stock market is analyzed by Shahzad et al., 2021a who examine the asymmetric volatility spillover among sectors using a VAR model. They highlight the sectors, such as industries, utilities, energies and materials, to which governments should pay more attention to maintain the stability of the Chinese stock market since they experience high volatility spillovers. From policy-making perspective, Bouri et al., 2021 analyze the changes in the connections of the returns of assets such as gold, crude oil, USD index, bond index, and MSCI World

index, during the epidemic. They show how the network is affected by the epidemic, in particular, MSCI World and USD index appear as primary transmitters of shock before the pandemic while the bond index appears primary transmitter during the Covid-19 period. These papers differ from our work in the choice of the modeling approach. The methodology used in the former is a VAR model to calculate the spillover asymmetry measure, while the latter apply a TVP-VAR by modeling the coefficients as a stochastic process. This work, instead applies the BGVAR model in Ahelegbey, Billio, and Casarin, 2016b **Ahelegbey2021Feb** to analyze the intra- and inter-layer connectivity between return indexes and sentiments of different industries.

In this paper, we propose a network model to stress the relationship among companies in different sectors, considering the dynamic pre and during Covid-19 pandemic. Our study focuses on the top 50 world companies due to their important role in the entire world economy and also due to the amount of available textual information. We want to assess how much the largest worldwide companies are suffering from the pandemic and whether the relationship between them is changing. To answer these questions we build a network model that considers two sources of information: textual data from news and blog and financial stock prices. We decided to analyze separately the different sectors to stress in which sector the pandemic affects more and how.

5.4 Methodology

5.4.1 Network VAR Model Formulation

Let R_t denote the returns of the stock market indices of n institutions at time t, and S_t denote the sentiment of the institutions. Let $Y_t = (R_t, S_t)$ be a $2n \times 1$ vector whose dynamic evolution can be described by a VAR(Y_t) process:

$$Y_t = \sum_{l=1}^{p} B_l \, Y_{t-s} + U_t \tag{5.1}$$

$$U_t = B_0 \, U_t + \varepsilon_t \tag{5.2}$$

where p is the lag order, B_l is $2n \times 2n$ matrix of coefficients with $B_{ij|l}$ measuring the effect of $Y_{j,t-l}$ on $Y_{i,t}$, U_t is a vector independent and identically normal residuals with covariance matrix Σ_u, B_0 is a zero diagonal matrix where $B_{ik}(0)$ records the contemporaneous effect of a shock to Y_k on Y_i, and ε_t is a vector of orthogonalized disturbances with covariance matrix Σ_ε. From (5.2), the Σ_u can be expressed in terms of B_0 and Σ_ε as

$$\Sigma_u \;=\; (I - B_0)^{-1} \Sigma_\varepsilon (I - B_0)^{-1'} \tag{5.3}$$

A network model is a convenient representation of the relationships among a set of variables. They are defined by nodes joined by a set of links, describing the statistical relationships between a pair of variables. The use of networks in VAR models helps to interpret the temporal and contemporaneous relationships in a multivariate time series. To analyze (5.1) and (5.2) through networks, we assign to each coefficient in B_l a corresponding latent indicator in $G_l \in \{0,1\}^{2n \times 2n}$, such that for $i, j = 1, \ldots, n$,

and $l = 0, 1, \ldots, p$:

$$B_{ij|l} = \begin{cases} 0 & \text{if } G_{ij|l} = 0 \implies Y_{j,t-l} \not\to Y_{i,t} \\ \beta_{ij} \in \mathbb{R} & \text{if } G_{ij|l} = 1 \implies Y_{j,t-l} \to Y_{i,t} \end{cases} \tag{5.4}$$

where $Y_{j,t-l} \not\to Y_{i,t}$ means that Y_j does not influence Y_i at lag l. We define two matrices $A \in \{0,1\}^{2n \times 2n}$ and $A^w \in \mathbb{R}^{2n \times 2n}$ such that

$$A = \mathbf{1}\left(\sum_{l=0}^{p}\sum_{ij} G_{ij|l} > 0\right) = \begin{pmatrix} A_{R|R} & A_{R|S} \\ A_{S|R} & A_{S|S} \end{pmatrix}, \quad A^w = \sum_{l=0}^{p}\sum_{ij} B_{ij|l} = \begin{pmatrix} A^w_{R|R} & A^w_{R|S} \\ A^w_{S|R} & A^w_{S|S} \end{pmatrix} \tag{5.5}$$

where $\mathbf{1}(G_{ij} > 0)$ is the indicator function, i.e., unity if $G_{ij} > 0$ and zero otherwise, $A^w_{R|R}$ and $A^w_{R|S}$ are sub-matrices of A^w that measure the cumulative effect of R_{t-l} and S_{t-l} on R_t for $l = 0, \ldots, p$, respectively. The sub-matrices of A reports the following:

$$A_{R|R}(i,j) = \begin{cases} 1, & \text{if } R_j \to R_i \\ 0, & \text{if } R_j \not\to R_i \end{cases}, \qquad A_{R|S}(i,k) = \begin{cases} 1, & \text{if } S_k \to R_i \\ 0, & \text{if } S_k \not\to R_i \end{cases} \tag{5.6}$$

$$A_{S|R}(q,j) = \begin{cases} 1, & \text{if } R_j \to S_q \\ 0, & \text{if } R_j \not\to S_q \end{cases}, \qquad A_{S|S}(q,k) = \begin{cases} 1, & \text{if } S_k \to S_q \\ 0, & \text{if } S_k \not\to S_q \end{cases} \tag{5.7}$$

where $R_j \to R_i$ exist if there is a contemporaneous or lagged directed link from R_j to R_i. Similar reasoning holds for $S_k \to R_i$, $R_j \to S_q$, and $S_k \to S_q$. Thus, $A_{R|R}$ specifies adjacency matrix of equity-to-equity connections, $A_{R|S}$ for sentiment-to-equity, $A_{S|R}$ for equity-to-sentiment, and $A_{S|S}$ for sentiment-to-sentiment linkages. $A^w = (A^w_{R|R}, A^w_{R|S}, A^w_{S|R}, A^w_{S|S})$ specifies the weights of the relationship in $A = (A_{R|R}, A_{R|S}, A_{S|R}, A_{S|S})$ obtained as a sum of the estimated contemporaneous and lagged coefficients. The correspondence between (G, B) and (A, A^w) is such that the former captures the short-run dynamics in $Y_t = (R_t, S_t)$ while the latter can be viewed as long-term direct relationships when $Y_t = Y_{t-1} = \ldots = Y_{t-p}$. Defining a sparse structure on (G, B) induces parsimony of the short-run model and sparsity on the long-run relationship matrices (A, A^w).

5.4.2 Bayesian Graphical Vector Autoregression

The model specification in (5.1) and (5.2) combines to form the structural VAR model which is well documented to exhibit identifiability problems. To circumvent this problem, we apply the Bayesian graphical vector autoregressive (BGVAR) approach of Ahelegbey, Billio, and Casarin, 2016a to separate and estimate the contemporaneous and lagged interactions associated with the VAR. We apply the BGVAR to study the intra- and inter-layer connectivity between return indexes and sentiments of different industries. We build on the collapsed Gibbs algorithm in Ahelegbey, Billio, and Casarin, 2016b **Ahelegbey2021Feb** by sampling the temporal dependence from its marginal distribution and the contemporaneous network from it's conditional distribution. We complete the Bayesian formulation with prior specification and posterior approximations to draw inference on the model parameters.

Prior Specification

We specify the prior distributions of $(p, G, B, \Sigma_\varepsilon)$ as follows:

$$p \sim \mathcal{U}(1, \bar{p}), \quad [B_{ij}|G_{ij} = 1] \sim \mathcal{N}(0, \eta), \quad G_{ij} \sim \text{Ber}(\pi_{ij}), \quad \Sigma_\varepsilon^{-1} \sim \mathcal{W}(\delta, \Lambda_0)$$

where $\bar{p}, \eta, \pi_{ij}, \delta$, and S_0 are hyper-parameters. The specification for p is a discrete uniform prior on the set $\{1, \ldots, \bar{p}\}$, $1 < \bar{p}$. The specification for B_{ij} conditional on G_{ij} follows a normal distribution with zero mean and variance η. Thus, relevant explanatory variables that predict a response variable must be associated with coefficients different from zero and the rest (representing not-relevant variables) are restricted to zero. We consider G_{ij} as Bernoulli distributed with π_{ij} as the prior probability. We assume Σ_ε^{-1} is Wishart distributed with prior expectation $\frac{1}{\delta}\Lambda_0$ and $\delta > n$ the degrees of freedom parameter.

Posterior Approximation

Let $X_t = (Y'_{t-1}, \ldots, Y'_{t-p})'$ be an $np \times 1$ vector of lagged observations, denote with $Y = (Y_1, \ldots, Y_N)$ a $N \times n$ matrix collection of all observations, and $X = (X_1, \ldots, X_N)$ be an $N \times np$ matrix collection of lagged observations. We fixed $p = 5$ to allow us select the relevant variables in different equations of the system. Following the Bayesian framework of Geiger and Heckerman, 2002, we integrate out the structural parameters analytically to obtain a marginal likelihood function over graphs. Following Ahelegbey and Giudici, 2020, we approximate inference of the parameters via a collapsed Gibbs sampler such that the algorithm proceeds as follows:

1. Sample via a Metropolis-within-Gibbs $[G_0, G_{1:p}|Y, p]$ by
   (a) Sampling from the marginal distribution: $[G_{1:p}|Y, p]$
   (b) Sampling from the conditional distribution: $[G_0|Y, p, G_{1:p}]$
2. Sample from $[B_0, B_{1:p}, \Sigma_\varepsilon|Y, \hat{G}_0, \hat{G}_{1:p}, p]$ by iterating the following steps:
   (a) Sample $[B_{i,\pi_i|1:p}|Y, \hat{G}_{1:p}, \hat{G}_0, B_0, \Sigma_\varepsilon] \sim \mathcal{N}(\hat{B}_{i,\pi_i|1:p},\ D_{\pi_i})$ where

$$\hat{B}_{i,\pi_i|1:p} = \sigma_{u,i}^{-2} D_{\pi_i} X'_{\pi_i} Y_i, \qquad D_{\pi_i} = (\eta^{-1} I_{d_x} + \sigma_{u,i}^{-2} X'_{\pi_i} X_{\pi_i})^{-1} \tag{5.8}$$

   where $X_{\pi_i} \in X$ corresponds to $(\hat{G}_{y_i, x_\pi|1:p} = 1)$, $\sigma^2_{u,i}$ is the i-th diagonal element of $\hat{\Sigma}_u = (I - \hat{B}_0)^{-1}\hat{\Sigma}_\varepsilon(I - \hat{B}_0)^{-1'}$, and d_x is the number of covariates in X_{π_i}.
   (b) Sample $[B_{i,\pi_i|0}|Y, \hat{G}_0, \hat{G}_{1:p}, B_{1:p}, \Sigma_\varepsilon] \sim \mathcal{N}(\hat{B}_{i,\pi_i|0}, Q_{\pi_i})$ where

$$\hat{B}_{i,\pi_i|0} = \sigma_{\varepsilon,i}^{-2} Q_{\pi_i} \hat{U}'_{\pi_i} \hat{U}_i, \qquad Q_{\pi_i} = (\eta^{-1} I_{d_u} + \sigma_{\varepsilon,i}^{-2} \hat{U}'_{\pi_i} \hat{U}_{\pi_i})^{-1} \tag{5.9}$$

   where $\hat{U} = Y - X\hat{B}'_{1:p}$, $\hat{U}_{\pi_i} \in \hat{U}_{-i}$ is the set of contemporaneous predictors of $\hat{U}_i$ that corresponds to $(\hat{G}_{y_i, y_\pi|0} = 1)$, and d_u is the number of covariates in U_{π_i}
   (c) Sample $[\Sigma_\varepsilon^{-1}|Y, \hat{G}_{1:p}, \hat{G}_0, B_{1:p}, B_0] \sim \mathcal{W}(\delta + N,\ \Lambda_N)$ where

$$\Lambda_N = \Lambda_0 + (\hat{U} - \hat{U}\hat{B}'_0)'(\hat{U} - \hat{U}\hat{B}'_0) \tag{5.10}$$

See Ahelegbey and Giudici, 2020 for a detailed description of the network sampling algorithm and convergence diagnostics.

For our empirical application, we set the hyper-parameters as follows: $\pi_{ij} = 0.5$ (which leads to a uniform prior on the graph space), $\eta = 100, \delta = n + 2$ and $\Lambda_0 = \delta I_n$. We set the number of MCMC iterations to sample 200,000 graphs and we ensured that the convergence and mixing of the MCMC chains are tested via the potential scale reduction factor (PSRF) of Gelman and Rubin, 1992.

5.5 Data

For our analysis, we focus on some of the most important American companies: the top 50 of the S&P. We obtain the daily stock prices of these companies from yahoo finance covering a period that ranges from August 2016 to November 17th, 2020. We also employ a sentiment index referred to the same companies and period produced by Brain[2] a research company .

No.	Stock	Ticker
	Communication	
1	AT & T Inc.	T
2	Verizon Comm. Inc.	VZ
	Consumer	
3	Amazon.com Inc.	AMZN
4	Comcast Corp.	CMCSA
5	Walt Disney Co.	DIS
6	Home Depot Inc.	HD
7	McDonald's Corp.	MCD
8	Netflix Inc.	NFLX
9	Costco Wholesale Corp.	COST
10	Coca-Cola Co.	KO
11	PepsiCo Inc.	PEP
12	Procter & Gamble Co.	PG
13	Philip Morris Int. Inc.	PM
14	Walmart Inc.	WMT
	Financial	
15	Bank of America Corp	BAC
16	Berkshire Hathaway Inc.	BRK.B
17	Citigroup Inc.	C
18	JPMorgan Chase & Co.	JPM
19	Wells Fargo & Co.	WFC
	Industrial	
20	Boeing Co.	BA
21	Honeywell Int. Inc.	HON
22	Union Pacific Corp.	UNP
23	Raytheon Technologies	RTX
	Energy	
24	Chevron Corp.	CVX
25	Exxon Mobil Corp.	XOM
	Health Care	
26	AbbVie Inc.	ABBV
27	Abbott Laboratories	ABT
28	Amgen Inc.	AMGN
29	Bristol-Myers Squibb Co.	BMY
30	Johnson & Johnson	JNJ
31	Medtronic Plc	MDT
32	Merck & Co. Inc.	MRK
33	Pfizer Inc.	PFE
34	Thermo Fisher Scientific Inc.	TMO
35	UnitedHealth Group Inc.	UNH
	Tech	
36	Apple Inc.	AAPL
37	Accenture Plc	ACN
38	Adobe Inc.	ADBE
39	Broadcom Inc.	AVGO
40	Salesforce.com inc.	CRM
41	Cisco Systems Inc.	CSCO
42	Facebook Inc.	FB
43	Alphabet Inc.	GOOGL
44	Intel Corp.	INTC
45	Mastercard Inc.	MA
46	Microsoft Corp.	MSFT
47	NVIDIA Corp.	NVDA
48	Oracle Corp.	ORCL
49	PayPal Holdings Inc.	PYPL
50	Visa Inc.	V

TABLE 5.1: Detailed description of the top 50 S&P companies.

Brain is a research company specialized in the production of alternative datasets and in the development of proprietary algorithms for investment strategies on financial markets. The Brain Sentiment Indicator dataset comprises of a daily sentiment indicator for the largest listed worldwide companies. Such indicator represents a

[2]https://braincompany.co

	Equity Returns					Sentiment Scores			
	Mean	Std	Skew	Kurt		Mean	Std	Skew	Kurt
AAPL	0.14	1.94	-0.38	7.76	S_AAPL	0.06	0.12	-0.62	1.43
ABBV	0.06	1.87	-1.12	16.06	S_ABBV	0.16	0.28	-0.85	0.78
ABT	0.09	1.60	-0.11	7.29	S_ABT	0.21	0.28	-0.88	0.79
ACN	0.08	1.61	0.00	9.24	S_ACN	0.27	0.26	-0.71	1.11
ADBE	0.14	2.08	-0.01	10.02	S_ADBE	0.18	0.27	-0.74	0.72
AMGN	0.04	1.67	0.13	6.62	S_AMGN	0.18	0.26	-0.48	0.34
AMZN	0.13	1.90	0.09	4.31	S_AMZN	0.14	0.27	-0.61	0.53
AVGO	0.09	2.29	-1.33	15.83	S_AVGO	0.14	0.29	-0.45	-0.04
BA	0.05	3.01	-0.61	19.53	S_BA	0.03	0.21	0.03	-0.28
BAC	0.06	2.19	-0.12	13.08	S_BAC	0.10	0.20	-0.32	0.48
BMY	0.02	1.67	-1.51	11.08	S_BMY	0.17	0.29	-0.77	0.30
BRK.B	0.04	1.40	-0.41	13.60	S_BRK.B	0.13	0.30	-0.51	-0.39
C	0.02	2.38	-0.81	16.76	S_C	0.08	0.27	-0.55	0.35
CMCSA	0.04	1.67	-0.11	6.16	S_CMCSA	0.14	0.26	-0.54	0.32
COST	0.09	1.37	-0.17	8.80	S_COST	0.13	0.28	-0.48	0.09
CRM	0.11	2.15	0.50	17.83	S_CRM	0.15	0.31	-0.76	0.61
CSCO	0.04	1.77	-0.57	10.44	S_CSCO	0.19	0.23	-0.74	0.95
CVX	0.00	2.13	-1.44	35.74	S_CVX	0.07	0.29	-0.49	-0.09
DIS	0.04	1.76	0.24	14.49	S_DIS	0.09	0.21	-0.40	1.05
FB	0.07	2.11	-1.20	14.38	S_FB	-0.07	0.16	0.16	-0.23
GOOGL	0.07	1.71	-0.43	6.85	S_GOOGL	0.04	0.15	-0.18	0.51
HD	0.07	1.70	-2.15	34.62	S_HD	0.16	0.25	-0.60	0.66
HON	0.07	1.64	-0.27	14.71	S_HON	0.20	0.27	-0.74	0.87
INTC	0.04	2.18	-0.85	18.17	S_INTC	0.11	0.19	-0.29	0.00
JNJ	0.03	1.31	-0.68	11.50	S_JNJ	0.13	0.28	-0.62	0.32
JPM	0.06	1.95	-0.12	16.96	S_JPM	0.08	0.24	-0.43	0.80
KO	0.03	1.31	-1.03	11.96	S_KO	0.13	0.25	-0.48	0.29
MA	0.12	1.89	0.03	11.77	S_MA	0.20	0.24	-0.87	1.21
MCD	0.07	1.51	-0.29	35.34	S_MCD	0.03	0.27	0.05	-0.10
MDT	0.03	1.60	-0.54	12.25	S_MDT	0.20	0.28	-0.39	0.08
MRK	0.04	1.41	-0.20	6.33	S_MRK	0.21	0.24	-0.65	0.61
MSFT	0.13	1.78	-0.36	12.06	S_MSFT	0.15	0.15	-0.56	2.27
NFLX	0.15	2.49	0.24	4.76	S_NFLX	0.09	0.18	-0.38	0.62
NVDA	0.21	3.10	-0.14	10.20	S_NVDA	0.16	0.21	-0.66	0.87
ORCL	0.04	1.68	0.49	23.15	S_ORCL	0.16	0.26	-0.86	1.08
PEP	0.04	1.37	-0.65	25.77	S_PEP	0.14	0.28	-0.56	0.19
PFE	0.02	1.43	-0.19	7.41	S_PFE	0.16	0.24	-0.63	0.75
PG	0.06	1.32	0.23	14.34	S_PG	0.14	0.31	-0.54	0.03
PM	0.00	1.74	-1.76	17.71	S_PM	0.07	0.32	-0.52	-0.30
PYPL	0.15	2.18	0.01	8.84	S_PYPL	0.15	0.27	-0.69	0.52
RTX	0.01	2.03	-0.38	14.71	S_RTX	0.14	0.30	-0.54	0.02
T	-0.01	1.53	-0.64	8.04	S_T	0.10	0.21	-0.24	0.09
TMO	0.11	1.61	-0.26	5.14	S_TMO	0.21	0.28	-0.91	1.13
UNH	0.09	1.89	-0.56	16.71	S_UNH	0.17	0.28	-0.55	0.30
UNP	0.08	1.80	-0.64	11.80	S_UNP	0.11	0.31	-0.40	-0.41
V	0.09	1.68	-0.22	13.19	S_V	0.15	0.27	-0.77	1.08
VZ	0.03	1.23	0.16	5.42	S_VZ	0.14	0.23	-0.48	0.52
WFC	-0.05	2.17	-0.49	12.61	S_WFC	-0.01	0.24	-0.02	0.40
WMT	0.08	1.41	0.70	15.61	S_WMT	0.08	0.23	-0.48	0.73
XOM	-0.06	1.83	-0.24	11.34	S_XOM	0.04	0.25	-0.15	-0.06

TABLE 5.2: Descriptive Statistics for Equity returns and Sentiment scores.

score that ranges between -1 and +1 and is based on financial news and blogs written in English. Each news is pre-analyzed to assign the corresponding company through the use of a dictionary of company names; then news are categorized using syntactic rules or machine learning classifiers. If this step fails a dictionary based approach is used.

The final dataset is composed of 1021 observations and 100 variables (for each company we have two columns: one for the closing market price and one for the

sentiment score). The complete list of companies is available in Table 5.1. Since the companies under analysis belong to different sectors, we have divided them into sub groups according to the S&P's division that considers 11 sectors: communication services, consumer discretionary, consumer staples, energy, financials, health care, industrials, materials, real estate, technology, utilities. Three sectors (namely materials, real estate and utilities) are not represented in our dataset, in addition we decided to unify the two consumer categories thus obtaining 7 final groups.

Table 5.2 reports the summary statistics of the first four moments (i.e., mean, standard deviation, skewness and excess kurtosis) of the equity returns and sentiment scores. The statistics show that almost all the equity returns and sentiment scores have a near-zero mean. The equity returns, however, appear more volatile than the sentiment scores. That is, the standard deviation of the equity returns are relatively higher (greater than 1) compared to that of the sentiment scores (less than 1). A greater majority of the returns and sentiments exhibit fairly symmetric behaviour, i.e., they are characterized mostly by small but consistent positive outcomes and, occasionally, large negative returns. The excess kurtosis of the sentiment scores are largely less than 3, which indicates that the sentiments data are approximately normal (via skewness-kurtosis summary), while that of the equity returns confirms the stylized facts of leptokurtic behavior of daily return series.

5.6 Results

We apply the BGVAR estimation methodology to study the dynamics of interconnectedness among the top 50 of S&P companies and the sub-sectors via a two-and-half year (approximately 504 days) rolling window. Our choice of window size is motivated by the need to have enough data points to capture 24-months dynamic dependence among the companies. We set the increments between successive rolling windows to one month. The first window covers August 2016 – July 2018, followed by September 2016 – August 2018, and the last from December 2018 – November 2020. In total, we have 29 rolling windows.

To unify the dataset, we compute the daily returns as the log difference of successive daily adjusted close prices of the companies equities. Since stock prices are not recorded for weekends, we consider the weekend sentiment scores in the computation of the Monday sentiment via a simple mean of the three days. In this way, both R_t and S_t express time variations: in the equity price and in the sentiment scores, respectively, for each company.

We study the equity-sentiment interconnectedness of the top 50 of S&P companies by considering them jointly as well as sub-sectors separately. Following the sector division of the companies in Table 5.1, we created five categories, namely: Consumer, Financial, Health Care, Tech and Miscellaneous and analyzed the interconnectedness for each sub-sector. Due to the low number of companies in the communication, energy and industrial sector, we combined them to create a unique sub-sector, which we refer to as "Miscellaneous".

We compare the sub-period networks, the pre-COVID-19 phase and the COVID-19 (Wave-1 and Wave-2) phase in terms of the number of links, the network density, the average degree, the clustering coefficient, and the average path length. We characterize, through numerical summaries, the time-varying nature of interconnections by monitoring the network density, average degree, clustering coefficient and average path length. In Figure 5.1, we report the evolution of the density of equity-sentiment interconnectedness along the analyzed period. Two curves are compared,

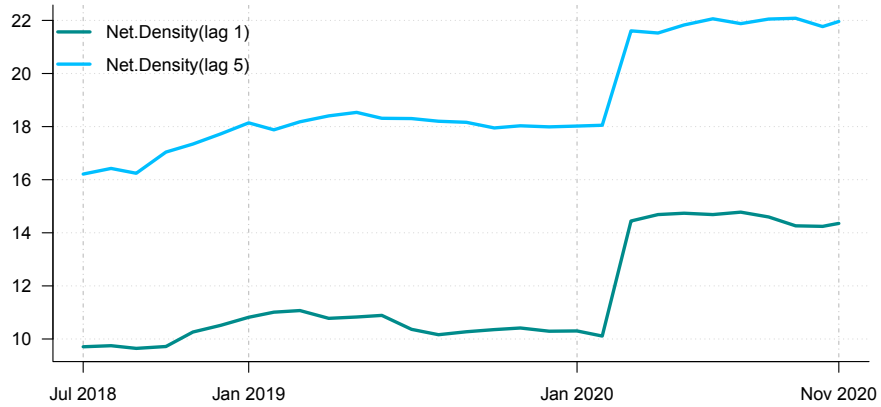
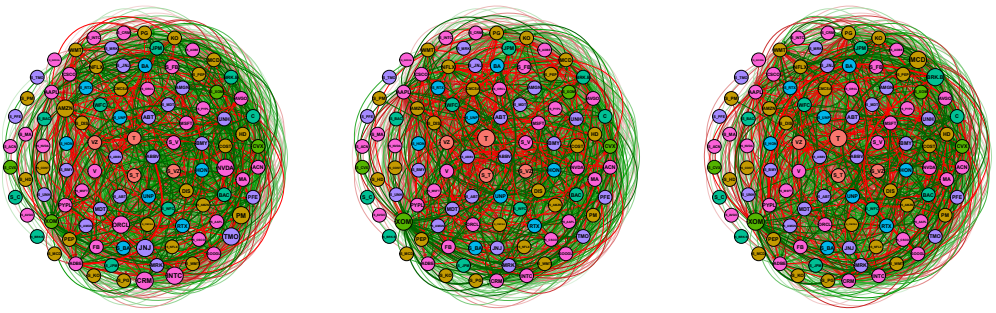


FIGURE 5.1: Density of equity-sentiment interconnectedness among top 50 S&P companies.

different in terms of employed lags, 1 vs 5. It clearly emerges the presence of two separated periods as of starting from late June 2020. The density increases by a factor of 6.5 – 8 points. As the pandemic enters in the most hitting phase, all the connections increase greatly, meaning that the exogenous shock affects the entire system as a whole, increasing the vulnerability as well. Indeed, a more interconnected system amplifies more and more any impact through a contagion spreading mechanism (see Cerchiello and Giudici, 2016c).

If we focus on the three periods of the data at hand (pre-pandemic, first wave, second wave), we can visualize the networks in Figure 5.2 and the metrics in Table 5.3 (you can visualize the networks only so to stress the difference with the equity-to-sentiments sub-period networks). If the difference between the pre-covid and the first wave is not so evident, we notice a change in the values of the number of links, the density, average degree and average path length during the second wave. This suggest that, although the system appears resilient as the first wave arrives, with the prolongation of the pandemic, companies can not stand any longer the shock.



(a) Pre-COVID-19 (b) COVID-19 Wave-1 (c) COVID-19 Wave-2

FIGURE 5.2: Sub-period network before and during COVID-19 period

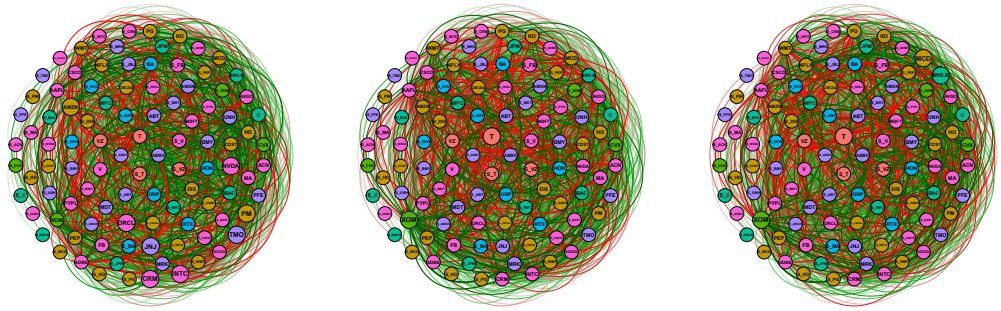
5.6.1 Equity-to-Equity Networks

To further analyze the Covid-19 pandemic effects on the system, we split the analysis in the two components: equity data on one hand and sentiment data on the other. In

Period	Links	Density	Average Degree	Clustering Coefficient	Average Path Length
Pre-COVID-19	2437	24.616	24.37	0.964	1.451
COVID-19 Wave-1	2401	24.253	24.01	0.994	1.131
COVID-19 Wave-2	2335	23.586	23.35	0.991	1.139

TABLE 5.3: The network statistics for sub-period interconnectedness before and during COVID-19 period.

particular we investigate what happens to the Equity to Equity connections, that is focusing only on the intra-equities layer linkages. As far as we are concerned with



(a) Pre-COVID-19 (b) COVID-19 Wave-1 (c) COVID-19 Wave-2

FIGURE 5.3: Equity-to-Equity sub-period network before and during COVID-19 period

Period	Links	Density	Average Degree	Clustering Coefficient	Average Path Length
Pre-COVID-19	2392	97.633	47.84	0.997	1.024
COVID-19 Wave-1	2395	97.755	47.90	0.999	1.022
COVID-19 Wave-2	2330	95.102	46.60	0.995	1.049

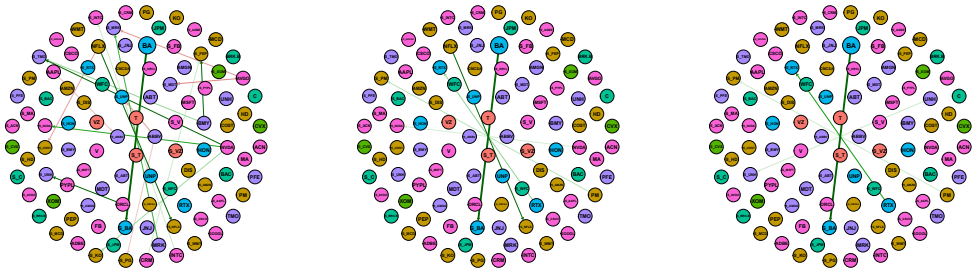
TABLE 5.4: Statistics for sub-period Equity-to-Equity interconnectedness before and during COVID-19 period.

the equity market, Figure 5.3 and Table 5.4 report the results. In particular, Table 5.4 discloses the pattern of the network along the three periods: similarly to previous results the financial market reacts more during the second wave. As we would have expected, there is a huge number of links which remains rather stable, confirming once again the deep interconnection of the financial market.

In Figure 5.4 and Table 5.5 we report another set of results looking at the effect of equity markets on sentiments, that is capturing sentiment reactions to changes in financial market performance. The reader can immediately notice a number of

Period	Links	Density	Average Degree
Pre-COVID-19	17	0.68	0.34
COVID-19 Wave-1	5	0.20	0.10
COVID-19 Wave-2	4	0.16	0.08

TABLE 5.5: Statistics for sub-period Equity-to-Sentiment network before and during COVID-19 period.



(a) Pre-COVID-19 (b) COVID-19 Wave-1 (c) COVID-19 Wave-2

FIGURE 5.4: Equity-to-Sentiments sub-period network before and during COVID-19 period

relevant facts: the total number of links is much less. That is to say that the financial market has a lower impact on sentiments and this is consistent along the whole time horizon. However, there is a clear difference in the three periods: the pre-covid period recorded three times higher sentiment reactions to changes in financial market indexes than in the Covid-19 periods. We could say that the influence from the financial world to the public perception one has been frozen by the virus, lowering down largely the influence channel.

Such apparently weird result can be explained by considering the type of shock affecting the system. The pandemic has a completely different nature, it is an exogenous diffused and pervasive shock which cannot be assimilated to other system perturbations like the financial ones. Modern populations and economies are not prepared or used to cope with so impacting restrictions and limitations of daily life. Results in Figure 5.4 and Table 5.5 and similarly in Figure 5.5 and Table 5.6 seem to suggest that the companies reacted by lowering down the interrelations, that is to say that the system was frozen and in attendance of the events. Companies became isolated entities waiting for a clearer evolution of the virus spread, blocking investments and activities planning and this is reflected in basically no correlations in either directions (equity to sentiment or sentiment to equity).

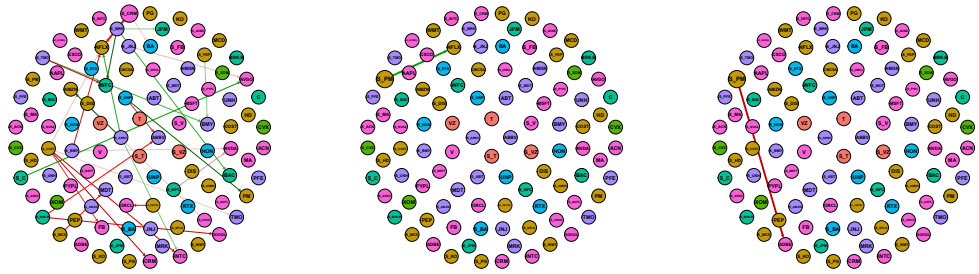
Such phenomenon is even more evident if we consider the opposite direction of transmission: from sentiment to equity. Figure 5.5 and Table 5.6 contain relative results and confirm the important dampening effect of the pandemic.

However, it is important to stress that such first analysis was run on the whole dataset, with no distinction made on the business sectors and looking at only one influence direction at the time (either equity to sentiment or sentiment to equity). That is to say that possible peculiar behaviours can occur sector by sector as we are going to discuss in the following. Just one connection survives during the first and

Period	Links	Density	Average Degree
Pre-COVID-19	28	1.12	0.56
COVID-19 Wave-1	1	0.04	0.02
COVID-19 Wave-2	1	0.04	0.02

TABLE 5.6: Statistics for sub-period Sentiment-to-Equity network before and during COVID-19 period.

the second wave. In the first wave, the only surviving linkage is $S\_PM \rightarrow NFLX$, and $S\_PM \rightarrow ADBE$ survived in the second wave. The reaction of Netflix and

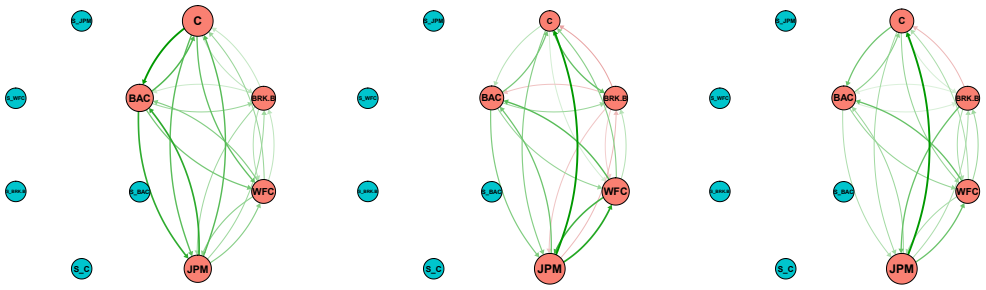


(a) Pre-COVID-19 (b) COVID-19 Wave-1 (c) COVID-19 Wave-2

FIGURE 5.5: Sentiment-to-Equity sub-period network before and during COVID-19 period

Adobe to sentiments associated with Philip Morris Int. - a tobacco company, are the only surviving linkages during the Covid pandemic.

Given the heterogeneity of the activities of the 50 companies at hand, it is relevant to deepen the analysis with regards to each specific sub-sector. Starting from the Financial sector, we notice from Figure 5.6 and Table 5.7 that all the indexes remains exactly the same. Our analysis reveals that the linkage among the finan-



(a) Pre-COVID-19 (b) COVID-19 Wave-1 (c) COVID-19 Wave-2

FIGURE 5.6: Sub-period Financial sub-sector network before and during COVID-19 period

Period	Links	Density	Average Degree	Clustering Coefficient	Average Path Length
Pre-COVID-19	20	22.222	2	1	1
COVID-19 Wave-1	20	22.222	2	1	1
COVID-19 Wave-2	20	22.222	2	1	1

TABLE 5.7: Statistics for sub-period Financial sub-sector network before and during COVID-19 period.

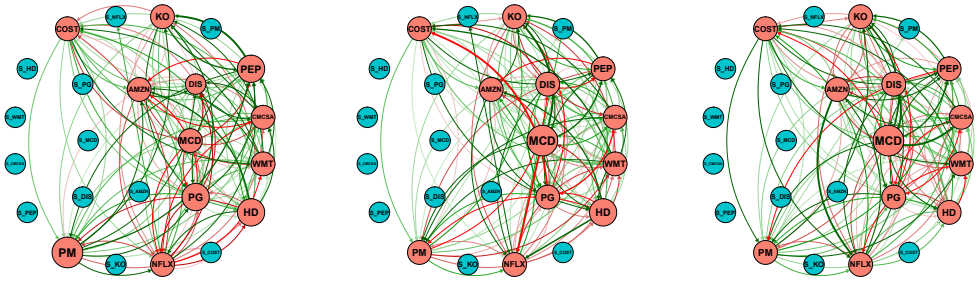
cial institutions revolve around their equity market performance with no effect from sentiments. Thus, the change in the networks structure that we have noticed in the previous tables, is not driven by the financial companies. We, however, notice that although the connections remain unchanged during the pre-covid and covid periods, the sign and magnitude of the interactions seems to change over the sub-periods. More specifically, Citigroup (C) and Berkshire Hathaway (BRK.B) seem to exhibit bi-directional relationship through out all periods. However, the pre-covid reported an almost equally positive link. The first and second wave of the Covid, however, recorded a positive impact of C on BRK.B, and a negative reverse impact

	Pre-COVID-19	COVID-19 Wave-1	COVID-19 Wave-2
	Top 5 Hub-Centrality Score		
1	C (0.579)	JPM (0.684)	JPM (0.758)
2	JPM (0.500)	WFC (0.496)	BAC (0.413)
3	BAC (0.484)	BAC (0.423)	C (0.339)
4	WFC (0.343)	BRK.B (0.275)	BRK.B (0.297)
5	BRK.B (0.250)	C (0.180)	WFC (0.228)
	Top 5 Authority-Centrality Score		
1	BAC (0.572)	C (0.656)	C (0.712)
2	JPM (0.480)	BAC (0.441)	WFC (0.550)
3	WFC (0.438)	WFC (0.435)	BAC (0.354)
4	C (0.407)	JPM (0.388)	JPM (0.233)
5	BRK.B (0.292)	BRK.B (0.187)	BRK.B (0.110)

TABLE 5.8: Hub and Authority Centrality of Financial sector network before and during COVID-19 period.

of BRK.B on C. A look at the centrality of the network in terms of Hub and Authority scores (in Table 5.8) shows that of the 5 companies, Citigroup was central to risk transmission during the pre-covid period, while JPM dominate in the Covid period.

In analyzing the Consumer sub-sector, Figure 5.7 shows the resulting network structure over the three sub-periods. Unlike, the Financial sub-sector, the Consumer sub-sector network record links are at all levels: equity-equity, equity-sentiment, sentiment-equity. For instance, the sentiment associated with Netflix (S_NFLX) react strongly positive to the equity performance of Netflix (NFLX) during the pre-covid, which reduced slightly in the first wave of the Covid but varnished in the second wave. We also observe a reaction from Netflix (NFLX) to the sentiment associated with Philip Morris Int. (S_PM). From Table 5.9, we notice a slight variation in the



(a) Pre-COVID-19 (b) COVID-19 Wave-1 (c) COVID-19 Wave-2

FIGURE 5.7: Sub-period Consumer sub-sector network before and during COVID-19 period

Period	Links	Density	Average Degree	Clustering Coefficient	Average Path Length
Pre-COVID-19	132	23.913	5.500	0.950	1.219
COVID-19 Wave-1	132	23.913	5.500	0.954	1.231
COVID-19 Wave-2	129	23.370	5.375	0.984	1.104

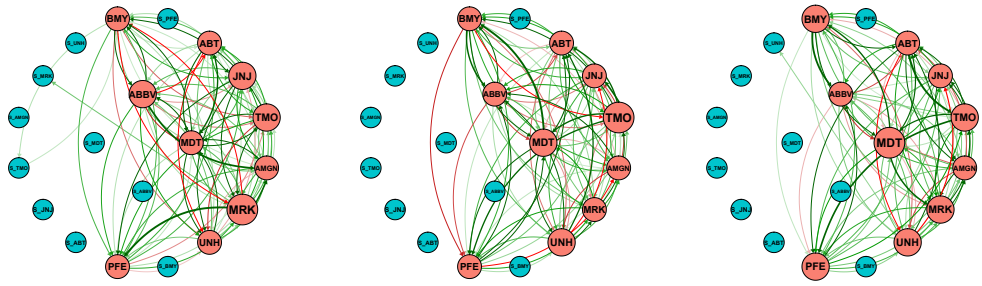
TABLE 5.9: Statistics for sub-period Consumer sub-sector network before and during COVID-19 period.

metrics of the second wave Consumer sub-sector network. In particular the clustering coefficient increases and the average path length decreases. Table 5.10 confirms the different behaviour of the consumer sub-sector: the hub companies during the pandemic change and increase in coefficient magnitude. The consumer system appears less resilient in comparison to the financial one. McDonalds, which is not in the top 5 hubs before the pandemics, not only appears all of a sudden, but it is also first ranked. Also Comcast Corp. and Amazon enter the ranking.

	Pre-COVID-19	COVID-19 Wave-1	COVID-19 Wave-2
	Top 5 Hub-Centrality Score		
1	PM (0.489)	MCD (0.578)	MCD (0.618)
2	PG (0.421)	HD (0.395)	PM (0.293)
3	PEP (0.343)	PG (0.339)	PEP (0.278)
4	HD (0.304)	CMCSA (0.266)	PG (0.276)
5	KO (0.279)	PEP (0.264)	AMZN (0.263)
	Top 5 Authority-Centrality Score		
1	PEP (0.400)	NFLX (0.436)	KO (0.454)
2	AMZN (0.384)	KO (0.385)	CMCSA (0.353)
3	NFLX (0.362)	AMZN (0.338)	PEP (0.345)
4	PG (0.349)	PM (0.289)	DIS (0.326)
5	KO (0.340)	CMCSA (0.288)	NFLX (0.310)

TABLE 5.10: Hub and Authority Centrality of Consumer sector network before and during COVID-19 period.

The Health-Care sub-sector network, represented in Figure 5.8 and Table 5.11, presents a pattern rather unstable. The indexes change without a common pattern, albeit showing an apparent drop in the magnitude during wave 1 and increasing again in wave 2. Similarly to consumer sub-sector, the links in Figure 5.8 are mixed



(a) Pre-COVID-19 (b) COVID-19 Wave-1 (c) COVID-19 Wave-2

FIGURE 5.8: Sub-period Health-Care sub-sector network before and during COVID-19 period

Period	Links	Density	Average Degree	Clustering Coefficient	Average Path Length
Pre-COVID-19	93	24.474	4.65	0.952	1.24
COVID-19 Wave-1	90	23.684	4.50	1.000	1.00
COVID-19 Wave-2	88	23.158	4.40	0.976	1.12

TABLE 5.11: Statistics for sub-period Health-Care sub-sector network before and during COVID-19 period.

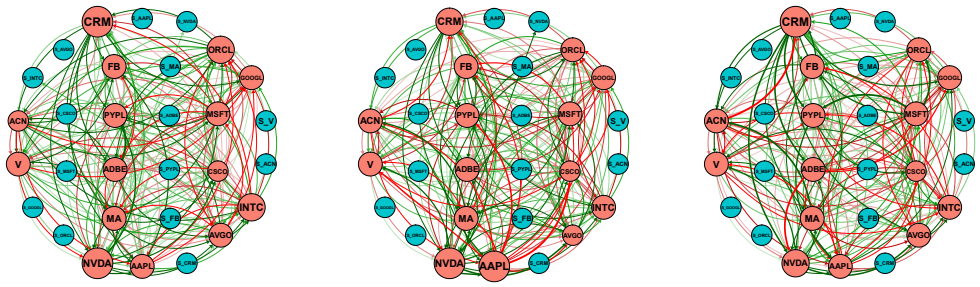
and both the hub and the authority indexes in Table 5.12 tend to change, not only

the rankings, but also the relevant companies. This suggest that the pandemic has deeply affected the health care sub-sector, as it is plausible to expect.

	Pre-COVID-19	COVID-19 Wave-1	COVID-19 Wave-2
	Top 5 Hub-Centrality Score		
1	MRK (0.495)	TMO (0.545)	MDT (0.455)
2	ABBV (0.407)	MDT (0.459)	TMO (0.399)
3	JNJ (0.394)	UNH (0.396)	PFE (0.334)
4	TMO (0.361)	MRK (0.266)	MRK (0.327)
5	UNH (0.273)	BMY (0.262)	UNH (0.302)
	Top 5 Authority-Centrality Score		
1	AMGN (0.465)	ABT (0.403)	ABT (0.378)
2	PFE (0.437)	BMY (0.392)	BMY (0.369)
3	BMY (0.356)	JNJ (0.367)	JNJ (0.344)
4	MDT (0.349)	PFE (0.343)	MDT (0.333)
5	MRK (0.276)	ABBV (0.325)	PFE (0.333)

TABLE 5.12: Hub and Authority scores of Health-Care sector network before and during COVID-19 period.

Figure 5.9 and Table 5.13 reports the network structure and its summary statistics for the Tech sector over the three sub-periods. What immediately emerges is the presence of much more connected networks regardless the period. The indexes are coherent and decrease as the periods pass by. Table 5.14 confirms the change in the



(a) Pre-COVID-19 (b) COVID-19 Wave-1 (c) COVID-19 Wave-2

FIGURE 5.9: Sub-period Tech sub-sector network before and during COVID-19 period

Period	Links	Density	Average Degree	Clustering Coefficient	Average Path Length
Pre-COVID-19	212	24.368	7.067	0.98	1.117
COVID-19 Wave-1	207	23.793	6.900	0.99	1.080
COVID-19 Wave-2	204	23.448	6.800	1.00	1.029

TABLE 5.13: Statistics for sub-period Tech sub-sector network before and during COVID-19 period.

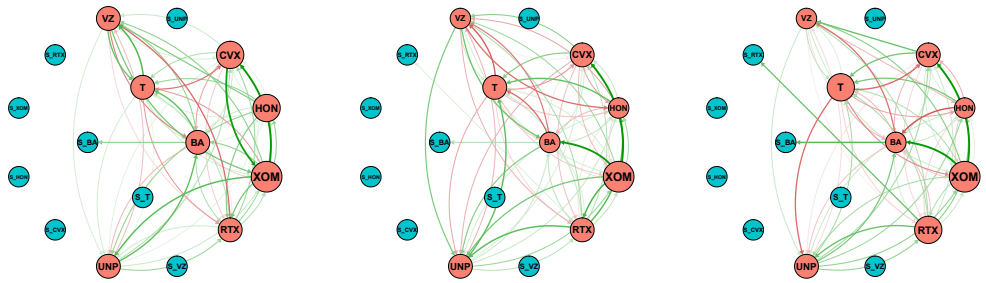
network structure: in particular two new players in the pandemic, namely Apple Inc. and Adobe Inc. for the hub score and Broadcom Inc. and Alphabet Inc. (Google) for the authority score.

The result of the miscellaneous sector which comprises Industrial, Communication and Energy companies are reported in Figure 5.10, Tables 5.15 and 5.16. We

	Pre-COVID-19	COVID-19 Wave-1	COVID-19 Wave-2
	Top 5 Hub-Centrality Score		
1	CRM (0.458)	AAPL (0.542)	CRM (0.511)
2	NVDA (0.453)	NVDA (0.541)	NVDA (0.390)
3	ORCL (0.371)	CRM (0.318)	ADBE (0.290)
4	INTC (0.280)	MSFT (0.236)	MSFT (0.289)
5	MSFT (0.265)	INTC (0.181)	AAPL (0.282)
	Top 5 Authority-Centrality Score		
1	PYPL (0.387)	PYPL (0.357)	AAPL (0.311)
2	ADBE (0.325)	MA (0.346)	PYPL (0.302)
3	MA (0.305)	CSCO (0.344)	V (0.295)
4	V (0.295)	V (0.329)	MA (0.277)
5	AAPL (0.288)	AVGO (0.318)	GOOGL (0.275)

TABLE 5.14: Hub and Authority Centrality of Tech sector network before and during COVID-19 period.

observe that similar to the Financial sub-sector, network among the group of companies in the miscellaneous sector is centered around the equity market performance, except for a links depicting the reaction of S_BAC (the sentiment associated with Bank of America) to the equity market performance of BAC (Bank of America). The



(a) Pre-COVID-19 (b) COVID-19 Wave-1 (c) COVID-19 Wave-2

FIGURE 5.10: Sub-period network of Miscellaneous sub-sector before and during COVID-19 period

Period	Links	Density	Average Degree	Clustering Coefficient	Average Path Length
Pre-COVID-19	56	23.333	3.500	0.960	1.125
COVID-19 Wave-1	58	24.167	3.625	0.923	1.194
COVID-19 Wave-2	56	23.333	3.500	0.923	1.222

TABLE 5.15: Statistics for sub-period Miscellaneous sub-sector network before and during COVID-19 period.

centrality ranking of the companies in this sub-sector shows that despite some slight changes in the top 5 companies, XOM (Exxon Mobile) and CVX (Chevron Corp.) remain the most central in terms of shock transmission and receiving risk, respectively, over the three sub-periods.

All the previous analysis can have important implications both for institutions and policy makers. It is well know in the literature ((Nicola, Cerchiello, and Aste, 2020b),(Cerchiello and Giudici, 2016c) (Cerchiello, Giudici, and Nicola, 2017)) that the interconnections among actors belonging to an economic system play a crucial

	Pre-COVID-19	COVID-19 Wave-1	COVID-19 Wave-2
	Top 5 Hub-Centrality Score		
1	XOM (0.567)	XOM (0.782)	XOM (0.680)
2	HON (0.377)	RTX (0.419)	T (0.471)
3	CVX (0.36)	T (0.284)	RTX (0.394)
4	UNP (0.354)	CVX (0.235)	UNP (0.289)
5	T (0.294)	UNP (0.194)	CVX (0.203)
	Top 5 Authority-Centrality Score		
1	CVX (0.578)	CVX (0.510)	CVX (0.555)
2	XOM (0.382)	UNP (0.425)	BA (0.459)
3	BA (0.334)	BA (0.422)	UNP (0.404)
4	VZ (0.325)	RTX (0.359)	VZ (0.324)
5	UNP (0.321)	HON (0.340)	HON (0.283)

TABLE 5.16: Centrality of Miscellaneous sectors network before and during COVID-19 period.

role during turbulence and crisis phases. A strongly interconnected network of companies can be considered either more resilient or vulnerable to shocks according to such events nature. In periods of financial crisis, arising from both real economy or financial markets, a high degree of interconnections can exert an extremely impacting systemic risk, which can cause a collapse of the whole sector given the strong dependencies among the actors. On the contrary, exogenous diffused shocks, not originally related to financial causes like the pandemic ones, can be much more impacting on poorly interconnected systems in which isolated companies can likely experience lack of aid from economic and sector peers increasing their probability of failure. Nevertheless, crisis not induced by real economy or equity markets can trigger very quickly consequent financial crisis, making even more difficult the evaluation of the optimal interconnections degree of a sector. That said, it is evident the reason why the monitoring and assessment of the levels of interconnection of the economic sectors should represent one of the main concern for regulators and supervisors. In case of downturns, it is important to quantify and monitor the level of vulnerability of the systems. To the same aim, the clear identification of pivotal actors (measured through hubs and authorities scores) can help in avoiding the activation of domino effects by supplying financial aids or cutting down connections. Moreover, the breakdown of the analysis in waves, is not only interesting from a descriptive point of view but it can represent a useful monitoring tool for non economically driven crisis like the pandemics. As the virus spreading moves on, the crisis becomes progressively and rapidly double-edged: epidemic and economic. Lastly, by directly leveraging also on the public sentiment, we account for the moods and perceptions of populations rather than only for speculators and investors. Such implementation allows for a more comprehensive and holistic view of the economic status, putting policy makers in an informed and aware framework.

5.7 Conclusions

The Covid-19 pandemic has deeply affected the population and all the relative activities. Health impact, social restrictions, economic downturn, overall instability are all direct consequences of the spread of the virus. Researchers worldwide have focused on studying, measuring and assessing such consequences at the different

levels. In this paper, we cope with the analysis of the economic impact of the pandemic, looking at the US top 50 companies of S&P market. In particular, we employ advanced network models able to leverage the temporal-dynamic dimension of the phenomenon through a novel specification of a Bayesian graphical vector autoregressive (BGVAR) approach. Moreover, we do not only rely on market data but emphasize the population perception and opinions by adding to the analysis a sentiment index built upon blogs and regular news. The analysis has revealed several interesting findings. First of all, the American financial market appears rather resilient as the first wave arrives but it is not able to stand the second one. The shock hits the whole system, increasing the interconnections and consequently the associated system risk. However the sub-sectors, which the 50 companies belong to, show different reactions, fully connected with the involved type of business. The Financial sector shows a particular resilience since all the indexes remains exactly the same. The linkage among the financial institutions revolve around their equity market performance with no effect from sentiments. Differently from the financial sector, the consumer one witnesses the strong interconnection between the equity and the sentiment components. Moreover, we notice clear signs of reactions as the pandemic moves on. The health-care sector is, as we would expect, affected by the instability induced by the pandemic. There is no a clear common pattern in the evolution of the networks, but it definitely reacts to the turbulence especially if we look at the most important hubs and authorities. Regarding the big Tech we obtain much more connected networks regardless the period. It is interesting to notice two new central players in the pandemic, namely Apple Inc. and Adobe Inc. for the hub score and Broadcom Inc. and Google for the authority score.

Moreover, we contribute to the ongoing discussions on the spillover effects of news and investors sentiment on equity returns in financial markets and interconnectedness among sectors. Some of our findings are as follows: from the equity-sentiment nexus there is evidence of more equity-to-sentiment pre-Covid than during the Covid-19 pandemic outbreak. There is more sentiment-to-equity pre-Covid-19 than during the Covid-19 pandemic outbreak, before Covid-19, more sentiment-to-equity influence than equity-to-sentiment. Finally, during the Covid-19 pandemic, more equity-to-sentiment than sentiment-to-equity influence. For what concerns the sectoral interconnectedness, we found that there is no significant difference in total interconnectedness among financial, consumer and miscellaneous sub-sectors. A drop in interconnectedness among health and tech sub-sectors, but with a much closely connected community and faster rate of shock propagation in COVID-19 Wave-1 and Wave-2, respectively.

With our work we contributed to recent literature aimed at analyzing the impact of Covid-19 pandemic. The main innovation is combining financial information and public opinions in an advanced network model exploiting the temporal dimension of the phenomenon by using rolling windows that reflect the market dynamics and the public perception shaping mechanism.

Further improvement of this study would consider up to date data, as the pandemic keeps on hitting the whole system. Indeed, the recent start of the vaccination campaign would be a further variable of interest that for sure would impact, not only the virus diffusion, but also the renovate confidence of the economic sectors and the population sentiment. Moreover, an analogous study with comparative purposes would be extremely useful on top 50 European companies.

Chapter 6

Conclusions

In this thesis, we analyzed the impact of textual analysis in the financial field and the possible advantages of this application.

Textual data is a type of data that falls into the broader category of alternative data; the rapid growth of these data, has led to a surge in applications of textual analysis in multiple fields including finance.

In the above chapters, we have seen how it is possible to integrate textual data with financial ones, highlighting the advantages that we can obtain from its implementation and results.

Alternative data, such as textual ones, are a type of unstructured data, widely used in areas such as politics, economics, sustainability, health. These data have joined the traditional ones and, if correctly analyzed and studied, they can be used to improve company performance or to make future predictions. As said in previous chapters, we will assist to unprecedented grow of alternative data and theirs applications in the near future. Combining those data with more traditional ones will improve the decision making process enabling business to make more accurate predictions about the future.

Going into the specific of the financial sector, textual data are obtained from the web (social media, blog, news) while traditional data are from financial ratios reports, companies' balance sheets, market data, etc.

Nowadays, alternative data have became more valuable due to the Covid-19 pandemic crisis. We believe that this period has highlighted aspects that were previously underestimated. First of all, web communication which was growing significantly pre-crisis, have had an unprecedented peak. People felt more in need to express their opinions, thoughts, and fears through social networks or blogs. Self-isolation and social interaction restrictions have led to an increase in time spent on the web, worldwide. Secondly, it is clear that we can no longer consider any type of events like an isolated entity, but as the subprime mortgage crisis and then the Covid-19 pandemic had taught us, we live in a world even more connected than it used to be. Globalization has created an interconnected environment where also actors, who are not directly connected, can influence the total outcome of the network. These two aspects are jointly highlighted in the chapters 4 and 5. In these chapters, the added value given by textual data is obtained through sentiment analysis. The thoughts and opinions extracted from news and blogs are used to investigate the relationships between the largest companies in the US market, with a purpose to investigate how those business are connected to each other and identify future development. In 4 we use an information-theoretic measure while in 5 we apply a network model taking in consideration the pandemic.

This work aims to demonstrate how textual data, if used in conjunction with traditional data, allows a broader understanding of economic phenomena, emerging trends, and relationships between companies.

Using those data can improve our understanding and knowledge about a specific event, but we also have to understand that alternative data will not substitute traditional data such as stock market values or balance sheet values, but instead they will enrich them.

The main conclusions we have drawn through the application of textual analysis in the financial field are summarized here.

In Chapter 3 we implement a credit scoring model through textual analysis. Textual data is obtained through text mining techniques such as categorization of the text of bank transactions. Using ad hoc dictionaries and distance measures, we can classify each account transaction into qualitative categories. These categories have been grouped into 5 macro-categories: non-essential goods (shopping, travel, living), essential goods (pharmacies, markets), financial services and utilities (bank, petrol station, telco companies), revenue (dividends and transfer), and salaries (wages and pensions). For each category, we have created two variables such as frequency and total amount spent by client for a specific category.

To highlight the value obtained through textual data, we compared the results from a model based only on financial variables, one based only on textual variables, and finally one that combines the two types of information.

The comparison is conducted by analyzing mis-classification error, the area under the curve AUC, and ROC curve. Thanks to this comparison we could see how using textual data has a positive impact on the credit scoring model. Specifically, the distribution of errors improve regarding the type I error which is the error that causes the most damage. This analysis could also be improved further using topic modeling instead of creating dictionaries.

In Chapter 4 we use a theoretical-informative measure called "transfer entropy", to highlight the causal relationships among the first 50 companies of the Standard & Poor during the period from November 2018 to November 2020. The companies are divided by sectors: Consumer discretionary, Consumer Staples, Energy, Healthcare, Tech, Financial Industrial and Communications. For analyzing the causal relationships we use typically financial information such as stock market prices and sentiments extracted from blogs and media. For each business we have two variables a financial and a sentiment one and we can highlight how casual relationships are more frequent for variables derived from the same data source, i.e prices with prices and sentiment with sentiment but with strong connections also concerning the links between prices and sentiment and sentiment and prices. An aspect highlighted by our analysis is that influences between companies are not limited by sector but there are also strong connections between business operating in different sectors. This could be explained since large companies tend to operate in many different businesses

Finally, the sector that most of all showed a strong causal signal between sentiment and prices is the technological one. This work could be implemented considering several companies, perhaps smaller or operating purely in the European market.

In Chapter 5 we study how the pandemic has impacted the largest business in the S&P in terms of their relationships.

To achieve this, we apply a network model called (BGVAR) Bayesian graphical vector auto regressive, based on stock market information and sentiment information extracted from blogs and news. We compare networks considering three periods such as the pre-COVID-19 phase, the Wave-1 and Wave-2 of COVID-19 phase in terms of number of links, network density, average degree, clustering coefficient,

and average path length. Our analysis has shown how the shock related to the pandemic obviously hit all sectors and increased the interconnections between companies. Among the considered business, the financial sector appears to be the most resilient and less dependent on sentiment, while the consumer market shows a strong connection between prices and opinions.

It is also possible to highlight how the financial sector was able to face the first wave but failed to manage the second one, while a sector like the healthcare failed to manage both waves. This analysis could be implemented considering the start-up period of the vaccination campaign and the sentiment derived from the campaign.

Overall, in this thesis, we have analyzed the added value of textual analysis in finance. We started with the explanation of alternative data seen as a broad category that includes textual data, and then we saw several applications in the financial field. The first application of textual data has been done on a credit scoring model. Our second application included two network models, the fist uses information measures and the second one relies on a Bayesian graphical vector auto regressive approach.

We can conclude that combining textual data with traditional one in the financial sector can lead to more accurate information and therefore more in-depth analysis. There is no doubt that typically financial data are essential and very important to understand the fundamentals of a business but also our analysis showed how alternative data will be able to empower the overall decision making process giving a new competitive advantage for business in facing future challenges.

The methodologies applied in this thesis are different and capture multiple aspects of the topic, future research can be conducted implementing further methodologies and focusing on different aspects.

Appendix A

Publications and Collaborations

A.1 Publications

The following peer-reviewed journals have published works related to this thesis:

- Chapter 3 has been published as: CERCHIELLO, Paola; SCARAMOZZINO, Roberta. On the Improvement of Default Forecast Through Textual Analysis. Frontiers Artif. Intell., 2020, 3: 16.
- Chapter 4 has been published as: SCARAMOZZINO, Roberta; CERCHIELLO, Paola; ASTE, Tomaso. Information theoretic causality detection between financial and sentiment data. Entropy, 2021, 23.5: 621.
- Chapter 5 is currently under publication process and it is available at SSRN as: AHELEGBEY, Daniel Felix; CERCHIELLO, Paola; SCARAMOZZINO, Roberta. Network Based Evidence of the Financial Impact of Covid-19 Pandemic. Available at SSRN 3780954, 2021.

A.2 Collaborations

The following partnerships contributed to the development of these Chapters:

- Chapter 3 Paola Cerchiello from University of Pavia
- Chapter 4 Paola Cerchiello from University of Pavia and Tomaso Aste from University College of London (UCL)
- Chapter 5 Paola Cerchiello and Daniel Felix Ahelegbey from University of Pavia

Bibliography

Abdou, Hussein A and John Pointon (2011). "Credit scoring, statistical techniques and evaluation criteria: a review of the literature". In: *Intelligent systems in accounting, finance and management* 18.2-3, pp. 59–88.

Acemoglu, Daron, Asuman Ozdaglar, and Alireza Tahbaz-Salehi (2015a). "Systemic risk and stability in financial networks". In: *American Economic Review* 105.2, pp. 564–608.

— (2015b). "Systemic Risk and Stability in Financial Networks". In: *American Economic Review* 105.2, pp. 564–608.

Acemoglu, Daron et al. (2012). "The network origins of aggregate fluctuations". In: *Econometrica* 80.5, pp. 1977–2016.

Aggarwal, Charu C and ChengXiang Zhai (2012). *Mining text data*. Springer Science & Business Media.

Ahelegbey, Daniel Felix, Monica Billio, and Roberto Casarin (2016a). "Bayesian Graphical Models for Structural Vector Autoregressive Processes". In: *Journal of Applied Econometrics* 31.2, pp. 357–386.

— (2016b). "Sparse graphical vector autoregression: a Bayesian approach". In: *Annals of Economics and Statistics/Annales d'Économie et de Statistique* 123/124, pp. 333–361.

Ahelegbey, Daniel Felix, Paola Cerchiello, and Roberta Scaramozzino (2021). "Network Based Evidence of the Financial Impact of Covid-19 Pandemic". In: *Available at SSRN 3780954*.

Ahelegbey, Daniel Felix and Paolo Giudici (2020). *NetVIX - A Network Volatility Index of Financial Markets*. SSRN 3693806 (accessed on November 7, 2020).

Ahmed, Shaghil et al. (2020). *The Impact of COVID-19 on Emerging Markets Economies' Financial Conditions*.

Algaba, Andres et al. (2020). "Econometrics Meets sentiment: An Overview of Methodology and Applications". In: *Journal of Economic Surveys*.

Allahyari, Mehdi et al. (2017). "A brief survey of text mining: Classification, clustering and extraction techniques". In: *arXiv preprint arXiv:1707.02919*.

Alloghani, Mohamed et al. (2020). "A systematic review on supervised and unsupervised machine learning algorithms for data science". In: *Supervised and Unsupervised Learning for Data Science*, pp. 3–21.

Aqlan, Ameen Abdullah Qaid, B Manjula, and R Lakshman Naik (2019). "A Study of Sentiment Analysis: Concepts, Techniques, and Challenges". In: *Proceedings of International Conference on Computational Intelligence and Data Engineering*. Springer, pp. 147–162.

Aste, Tomaso (2019). "Cryptocurrency Market Structure: Connecting Emotions and Economics". In: *Digital Finance* 1.1-4, pp. 5–21.

Aste, Tomaso, William Shaw, and Tiziana Di Matteo (2010). "Correlation structure and dynamics in volatile markets". In: *New Journal of Physics* 12.8, p. 085009.

Baek, Seung Ki et al. (2005). "Transfer entropy analysis of the stock market". In: *arXiv preprint physics/0509014*.

Baker, Scott R et al. (2020). *The unprecedented stock market impact of COVID-19*.

Battiston, Stefano et al. (2012). "Liaisons Dangereuses: Increasing Connectivity, Risk Sharing, and Systemic Risk". In: *Journal of Economic Dynamics and Control* 36.8, pp. 1121–1141.

Billio, Monica et al. (2012). "Econometric measures of connectedness and systemic risk in the finance and insurance sectors". In: *Journal of Financial Economics* 104.3, pp. 535–559.

Blondel, Vincent D et al. (2008). "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008.

Boccaletti, Stefano et al. (2006). "Complex networks: Structure and dynamics". In: *Physics Reports* 424.4-5, pp. 175–308.

Bollen, Johan, Huina Mao, and Xiaojun Zeng (2011a). "Twitter mood predicts the stock market". In: *Journal of Computational Science* 2.1, pp. 1–8.

— (2011b). "Twitter mood predicts the stock market". In: *Journal of Computational Science* 2.1, pp. 1–8. ISSN: 1877-7503.

Bolton, Christine et al. (2010). "Logistic regression and its application in credit scoring". PhD thesis. University of Pretoria.

Bordino, Ilaria et al. (2012). "Web Search Queries Can Predict Stock Market Volumes". In: *PLoS One* 7.7, e40014. ISSN: 1932-6203.

Bouri, Elie et al. (2018). "Does global fear predict fear in BRICS stock markets? Evidence from a Bayesian Graphical Structural VAR model". In: *Emerging Markets Review* 34, pp. 124–142.

Bouri, Elie et al. (2021). "Return connectedness across asset classes around the COVID-19 outbreak". In: *International review of financial analysis* 73, p. 101646.

Caccioli, Fabio, Paolo Barucca, and Teruyoshi Kobayashi (2018). "Network models of financial systemic risk: a review". In: *Journal of Computational Social Science* 1.1, pp. 81–114.

Carvalho, Jonnathan, Adriana Prado, and Alexandre Plastino (2014). "A statistical and evolutionary approach to sentiment analysis". In: *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. Vol. 2. IEEE, pp. 110–117.

Cerchiello, P. and P. Giudici (2016a). "Conditional graphical models for systemic risk estimation". In: *Expert Syst. Appl.* 43, pp. 165–174. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2015.08.047.

Cerchiello, P. and G. Nicola (2018). "Assessing news contagion in finance". In: *Econometrics* 6.1, p. 5. ISSN: 2225-1146. DOI: 10.3390/econometrics6010005.

Cerchiello, Paola and Paolo Giudici (2016b). "Big data analysis for financial risk management". In: *Journal of Big Data* 3.1, pp. 1–12.

— (2016c). "Conditional Graphical models for systemic risk estimation". In: *Expert Systems with Applications* 43, pp. 165–174.

— (2016d). "How to measure the quality of financial tweets". In: *Quality & Quantity* 50.4, pp. 1695–1713. ISSN: 1573-7845.

Cerchiello, Paola, Paolo Giudici, and Giancarlo Nicola (2017). "Twitter data models for bank risk contagion". In: *Neurocomputing* 264, pp. 50–56. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2016.10.101.

Chan, Samuel WK and James Franklin (2011). "A text-based decision support system for financial sequence prediction". In: *Decision Support Systems* 52.1, pp. 189–198.

Chen, Conghui, Lanlan Liu, and Ningru Zhao (2020). "Fear Sentiment, Uncertainty, and Bitcoin Price Dynamics: The Case of COVID-19". In: *Emerging Markets Finance and Trade* 56.10, pp. 2298–2309.

Choi, Hyunyoung and Hal Varian (2012). "Predicting the Present with Google Trends". In: *Economic Record* 88.s1, pp. 2–9. ISSN: 0013-0249.

Colladon, A Fronzetti et al. (2020). *Forecasting Financial Markets with Semantic Network Analysis in the COVID-19 Crisis*. arXiv preprint arXiv:2009.04975.

Cornée, Simon (2019). "The relevance of soft information for predicting small business credit default: Evidence from a social bank". In: *Journal of Small Business Management* 57.3, pp. 699–719.

Costola, Michele et al. (2020). *Machine Learning Sentiment Analysis, Covid-19 News and Stock Market Reactions*. SAFE, Working Paper.

Cover, Thomas M (1999). *Elements of information theory*. John Wiley & Sons.

Cowie, Jim and Wendy Lehnert (1996). "Information extraction". In: *Communications of the ACM* 39.1, pp. 80–91.

Dalal, Mita K and Mukesh A Zaveri (2011). "Automatic text classification: a technical review". In: *International Journal of Computer Applications* 28.2, pp. 37–40.

De Bandt, Olivier and Philipp Hartmann (2000). "Systemic risk: a survey". In: *Available at SSRN 258430*.

Derouiche, Karim and Marius Frunza (2020). *How Did COVID-19 Shaped the Tweets Sentiment Impact upon Stock Prices of Sport Companies?* Available at SSRN 3649726.

Devika, MD, C[a] Sunitha, and Amal Ganesh (2016). "Sentiment analysis: a comparative study on different approaches". In: *Procedia Computer Science* 87, pp. 44–49.

Diebold, F. and K Yilmaz (2014). "On the Network Topology of Variance Decompositions: Measuring the Connectedness of Financial Firms". In: *Journal of Econometrics* 182.1, pp. 119–134.

Dimpfl, Thomas and Franziska Julia Peter (2013). "Using transfer entropy to measure information flows between financial markets". In: *Studies in Nonlinear Dynamics & Econometrics* 17.1, pp. 85–102.

Djeundje, Viani B et al. (2021). "Enhancing credit scoring with alternative data". In: *Expert Systems with Applications* 163, p. 113766.

Dörre, Jochen, Peter Gerstl, and Roland Seiffert (1999). "Text mining: finding nuggets in mountains of textual data". In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 398–401.

Eberendu, Adanma Cecilia et al. (2016). "Unstructured Data: an overview of the data of Big Data". In: *International Journal of Computer Trends and Technology* 38.1, pp. 46–50.

Eddy, Yosi Lizar and Engku Muhammad Nazri Engku Abu Bakar (2017). "Credit scoring models: Techniques and issues". In: *Journal of advanced research in business and management studies* 7.2, pp. 29–41.

Eisenberg, Larry and Thomas H Noe (2001). "Systemic risk in financial systems". In: *Management Science* 47.2, pp. 236–249.

Elliott, Matthew, Benjamin Golub, and Matthew O Jackson (2014). "Financial Networks and Contagion". In: *American Economic Review* 104.10, pp. 3115–3153.

Engelberg, Joseph (2008). "Costly information processing: Evidence from earnings announcements". In: *AFA 2009 San Francisco meetings paper*.

Faloutsos, Christos and Douglas W Oard (1998). *A survey of information retrieval and filtering methods*. Tech. rep.

Farahani, Farzad V, Waldemar Karwowski, and Nichole R Lighthall (2019). "Application of graph theory for identifying connectivity patterns in human brain networks: a systematic review". In: *frontiers in Neuroscience* 13, p. 585.

Feinstein, Jeffrey (2013). "Alternative data and fair lending". In: *LexisNexis*.

Feldman, Ronen (2013a). "Techniques and applications for sentiment analysis". In: *Communications of the ACM* 56.4, pp. 82–89.

— (2013b). "Techniques and applications for sentiment analysis". In: *Communications of the ACM* 56.4, pp. 82–89. ISSN: 0001-0782.

Feldman, Ronen and Ido Dagan (1995). "Knowledge Discovery in Textual Databases (KDT)." In: *KDD*. Vol. 95, pp. 112–117.

Flach, Peter A and Nicolas Lachiche (2004). "Naive Bayesian classification of structured data". In: *Machine learning* 57.3, pp. 233–269.

Fortunato, Santo (2010). "Community detection in graphs". In: *Physics Reports* 486.3-5, pp. 75–174.

Freeman, Linton C (1978). "Centrality in social networks conceptual clarification". In: *Social networks* 1.3, pp. 215–239.

Geiger, Dan and David Heckerman (2002). "Parameter Priors for Directed Acyclic Graphical Models and the Characterization of Several Probability Distributions". In: *Annals of Statistics* 30.5, pp. 1412–1440.

Gelman, A. and D. B. Rubin (1992). "Inference from Iterative Simulation Using Multiple Sequences, (with discussion)". In: *Statistical Science* 7, pp. 457–511.

Ghiassi, Manoochehr, James Skinner, and David Zimbra (2013). "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network". In: *Expert Systems with applications* 40.16, pp. 6266–6282.

Girvan, Michelle and Mark EJ Newman (2002). "Community structure in social and biological networks". In: *Proceedings of the national academy of sciences* 99.12, pp. 7821–7826.

Giudici, Paolo and Iman Abu-Hashish (2019). "What determines bitcoin exchange prices? A network VAR approach". In: *Finance Research Letters* 28, pp. 309–318.

Giudici, Paolo, Branka Hadji-Misheva, and Alessandro Spelta (2020). "Network based credit risk models". In: *Qual. Eng.* 32.2, pp. 199–211. ISSN: 0898-2112. DOI: 10.1080/08982112.2019.1655159.

Giudici, Paolo and Alessandro Spelta (2016). "Graphical network models for international financial flows". In: *Journal of Business & Economic Statistics* 34.1, pp. 128–138.

Gordini, Niccolo (2014). "A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy". In: *Expert systems with applications* 41.14, pp. 6433–6445.

Gormsen, Niels Joachim and Ralph SJ Koijen (2020). *Coronavirus: Impact on stock prices and growth expectations*.

Granger, Clive WJ (1969). "Investigating causal relations by econometric models and cross-spectral methods". In: *Econometrica: Journal of the Econometric Society*, pp. 424–438.

Groth, Sven S and Jan Muntermann (2011). "An intraday market risk management approach based on textual analysis". In: *Decision Support Systems* 50.4, pp. 680–691.

Grunert, Jens, Lars Norden, and Martin Weber (2005). "The role of non-financial factors in internal credit ratings". In: *Journal of Banking & Finance* 29.2, pp. 509–531.

Guizzo, Erico Marui (2003). "The essential message: Claude Shannon and the making of information theory". PhD thesis. Massachusetts Institute of Technology.

Guo, Guangming et al. (2016). "Personal credit profiling via latent user behavior dimensions on social media". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pp. 130–142.

Gupta, Vishal, Gurpreet S Lehal, et al. (2009). "A survey of text mining techniques and applications". In: *Journal of emerging technologies in web intelligence* 1.1, pp. 60–76.

Hand, David J and William E Henley (1997). "Statistical classification methods in consumer credit scoring: a review". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 160.3, pp. 523–541.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hennig-Thurau, Thorsten et al. (2004). "Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet?" In: *Journal of interactive marketing* 18.1, pp. 38–52.

Huang, Allen H, Amy Y Zang, and Rong Zheng (2014). "Evidence on the information content of text in analyst reports". In: *The Accounting Review* 89.6, pp. 2151–2180.

Iyer, Rajkamal et al. (2016). "Screening peers softly: Inferring the quality of small borrowers". In: *Management Science* 62.6, pp. 1554–1577.

Jiang, Cuiqing et al. (2018). "Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending". In: *Annals of Operations Research* 266.1, pp. 511–529.

Jivani, Anjali Ganesh et al. (2011). "A comparative study of stemming algorithms". In: *Int. J. Comp. Tech. Appl* 2.6, pp. 1930–1938.

Joshi, Kalyani, Bharathi N, and Jyothi Rao (June 2016). "Stock Trend Prediction Using News Sentiment Analysis". In: *International Journal of Computer Science and Information Technology* 8, pp. 67–76.

Jothimani, Dhanya, Ravi Shankar, and Surendra S Yadav (2018). "A big data analytical framework for portfolio optimization". In: *arXiv preprint arXiv:1811.07188*.

Kamalloo, Ehsan and Mohammad Saniee Abadeh (2010). "An artificial immune system for extracting fuzzy rules in credit scoring". In: *IEEE Congress on Evolutionary Computation*. IEEE, pp. 1–8.

Kearney, Colm and Sha Liu (2014). "Textual sentiment in finance: A survey of methods and models". In: *International Review of Financial Analysis* 33, pp. 171–185.

Keskin, Z and T Aste (2019). "Information-theoretic measures for non-linear causality detection: application to social media sentiment and cryptocurrency prices". In: *arXiv preprint arXiv:1906.05740*.

Khemais, Zaghdoudi, Djebali Nesrine, Mezni Mohamed, et al. (2016). "Credit scoring and default risk prediction: A comparative study between discriminant analysis & logistic regression". In: *International Journal of Economics and Finance* 8.4, p. 39.

Kok, Stanley and Pedro Domingos (2005). "Learning the structure of Markov logic networks". In: *Proceedings of the 22nd international conference on Machine learning*, pp. 441–448.

Kolchyna, Olga et al. (2015). "Twitter sentiment analysis: Lexicon method, machine learning method and their combination". In: *arXiv preprint arXiv:1507.00955*.

Kou, Gang et al. (2019). "Machine learning methods for systemic risk analysis in financial sectors". In: *Technological and Economic Development of Economy* 25.5, pp. 716–742.

Krieg, Mark L (2001). "A tutorial on Bayesian belief networks". In:

Krzanowski, Wojtek J and David J Hand (2009). *ROC curves for continuous data*. Crc Press.

Kulin, Merima et al. (2021). "A survey on machine learning-based performance improvement of wireless networks: PHY, MAC and network layer". In: *Electronics* 10.3, p. 318.

Kumar, B Shravan and Vadlamani Ravi (2016). "A survey of the applications of text mining in financial domain". In: *Knowledge-Based Systems* 114, pp. 128–147.

Kwon, Okyu and Jae-Suk Yang (2008). "Information flow between composite stock index and individual stocks". In: *Physica A: Statistical Mechanics and its Applications* 387.12, pp. 2851–2856.

Laney, Doug et al. (2001). "3D data management: Controlling data volume, velocity and variety". In: *META group research note* 6.70, p. 1.

Larsen, Vegard H and Leif A Thorsrud (2019). "The Value of News for Economic Developments". In: *Journal of Econometrics* 210.1, pp. 203–218.

Lee, Hee Soo (2020). "Exploring the Initial Impact of COVID-19 Sentiment on US Stock Market Using Big Data". In: *Sustainability* 12.16, p. 6648.

Levenshtein, Vladimir I (1966). "Binary codes capable of correcting deletions, insertions, and reversals". In: *Soviet physics doklady*. Vol. 10. 8. Soviet Union, pp. 707–710.

Liberti, José María and Mitchell A Petersen (2019). "Information: Hard and soft". In: *Review of Corporate Finance Studies* 8.1, pp. 1–41.

Loughran, Tim and Bill McDonald (2011). "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks". In: *The Journal of finance* 66.1, pp. 35–65.

— (2016). "Textual Analysis in Accounting and Finance: A survey". In: *Journal of Accounting Research* 54.4, pp. 1187–1230.

MacKinlay, A Craig (1997). "Event studies in economics and finance". In: *Journal of Economic Literature* 35.1, pp. 13–39.

Mamaysky, Harry (2020). *Financial Markets and News about the Coronavirus*. Available at SSRN 3565597.

Man, Xiliu, Tong Luo, and Jianwu Lin (2019). "Financial sentiment analysis (fsa): A survey". In: *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS)*. IEEE, pp. 617–622.

Manning, Christopher and Hinrich Schutze (1999). *Foundations of statistical natural language processing*. MIT press.

Mantegna, Rosario N (1999). "Hierarchical structure in financial markets". In: *The European Physical Journal B-Condensed Matter and Complex Systems* 11.1, pp. 193–197.

Marques, AI, Vicente García, and José Salvador Sánchez (2013). "A literature review on the application of evolutionary computing to credit scoring". In: *Journal of the Operational Research Society* 64.9, pp. 1384–1399.

Marschinski, Robert and Holger Kantz (2002). "Analysing the information flow between financial time series". In: *The European Physical Journal B-Condensed Matter and Complex Systems* 30.2, pp. 275–281.

Mayer-Schoenberger, Viktor and K Cukier (2013). "The rise of big data: how it's changing the way we think about the world". In: *Foreign affairs* 92, pp. 28–40.

Medhat, Walaa, Ahmed Hassan, and Hoda Korashy (2014). "Sentiment analysis algorithms and applications: A survey". In: *Ain Shams engineering journal* 5.4, pp. 1093–1113.

Mishev, Kostadin et al. (2020). "Evaluation of sentiment analysis in finance: from lexicons to transformers". In: *IEEE Access* 8, pp. 131662–131682.

Mpofu, Thabiso Peter and Macdonald Mukosera (2014). "Credit scoring techniques: a survey". In: *International Journal of Science and Research (IJSR)*, pp. 2319–7064.

Müller, Vincent C and Nick Bostrom (2016). "Future progress in artificial intelligence: A survey of expert opinion". In: *Fundamental issues of artificial intelligence*. Springer, pp. 555–572.

Netzer, Oded, Alain Lemaire, and Michal Herzenstein (2019). "When words sweat: Identifying signals for loan default in the text of loan applications". In: *Journal of Marketing Research* 56.6, pp. 960–980.

Newman, Mark EJ (2003). "The structure and function of complex networks". In: *SIAM review* 45.2, pp. 167–256.

Newman, Mark EJ and Michelle Girvan (2004). "Finding and evaluating community structure in networks". In: *Physical review E* 69.2, p. 026113.

Nicola, Giancarlo, Paola Cerchiello, and Tomaso Aste (2020a). "Information network modeling for US banking systemic risk". In: *Entropy* 22.11, p. 1331.

— (2020b). "Information Network Modeling for U.S. Banking Systemic Risk". In: *Entropy* 22.11, p. 1331. ISSN: 1099-4300. DOI: 10.3390/e22111331.

Nicola, Maria et al. (2020). "The socio-economic implications of the coronavirus pandemic (COVID-19): A review". In: *International Journal of Surgery (London, England)* 78, p. 185.

Nigam, Kamal et al. (2000). "Text classification from labeled and unlabeled documents using EM". In: *Machine learning* 39.2, pp. 103–134.

Niu, Beibei, Jinzheng Ren, and Xiaotao Li (2019). "Credit scoring using machine learning by combing social network information: Evidence from peer-to-peer lending". In: *Information* 10.12, p. 397.

Nyman, Rickard, Sujit Kapadia, and David Tuckett (2021). "News and narratives in financial systems: exploiting big data for systemic risk assessment". In: *Journal of Economic Dynamics and Control*, p. 104119.

Onnela, J-P, Kimmo Kaski, and Janos Kertész (2004). "Clustering and information in correlation based financial networks". In: *The European Physical Journal B* 38.2, pp. 353–362.

Pantaleo, Ester et al. (2011). "When do improved covariance matrix estimators enhance portfolio optimization? An empirical comparative study of nine estimators". In: *Quantitative Finance* 11.7, pp. 1067–1080.

Pedro, Jose San, Davide Proserpio, and Nuria Oliver (2015). "MobiScore: towards universal credit scoring from mobile phone data". In: *international conference on user modeling, adaptation, and personalization*. Springer, pp. 195–207.

Peralta, Gustavo and Abalfazl Zareei (2016). "A network approach to portfolio selection". In: *Journal of Empirical Finance* 38, pp. 157–180.

Pereira, Jose Manuel, Mario Basto, and Amelia Ferreira da Silva (2016). "The logistic lasso and ridge regression in predicting corporate failure". In: *Procedia Economics and Finance* 39, pp. 634–641.

Powers, David MW (2020). "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation". In: *arXiv preprint arXiv:2010.16061*.

Pozzi, Francesco et al. (2008). "Dynamical correlations in financial systems". In: *Complex Systems II*. Vol. 6802. International Society for Optics and Photonics, 68021E.

Provost, Foster and Tom Fawcett (2013). "Data science and its relationship to big data and data-driven decision making". In: *Big data* 1.1, pp. 51–59.

Qiu, Junfei et al. (2016). "A survey of machine learning for big data processing". In: *EURASIP Journal on Advances in Signal Processing* 2016.1, pp. 1–16.

Rajput, Nikhil Kumar, Bhavya Ahuja Grover, and Vipin Kumar Rathi (2020). *Word frequency and sentiment analysis of twitter messages during Coronavirus pandemic*. arXiv preprint arXiv:2004.03925.

Ranco, Gabriele et al. (2015). "The effects of Twitter sentiment on stock price returns". In: *PloS one* 10.9, e0138441.

Rao, Marepalli B and Calyampudi Radhakrishna Rao (2014). *Computational Statistics with R*. Elsevier.

Rao, Tushar, Saket Srivastava, et al. (2012). "Analyzing stock market movements using twitter sentiment analysis". In:

Ravi, Kumar and Vadlamani Ravi (2015). "A survey on opinion mining and sentiment analysis: tasks, approaches and applications". In: *Knowledge-based systems* 89, pp. 14–46.

Schaeffer, Satu Elisa (2007). "Graph clustering". In: *Computer science review* 1.1, pp. 27–64.

Schreiber, Thomas (2000). "Measuring information transfer". In: *Physical Review Letters* 85.2, p. 461.

Segall, Richard S, Qingyu Zhang, and Mei Cao (2009). "Web-based text mining of hotel customer comments using SAS® text miner and megaputer polyanalyst®". In: *SWDSI 2009*, pp. 141–152.

Shahzad, Syed Jawad Hussain, Thi Hong Van Hoang, and Elie Bouri (2021). "From pandemic to systemic risk: contagion in the US tourism sector". In: *Current Issues in Tourism*, pp. 1–7.

Shahzad, Syed Jawad Hussain et al. (2021a). "Asymmetric volatility spillover among Chinese sectors during COVID-19". In: *International Review of Financial Analysis* 75, p. 101754.

Shahzad, Syed Jawad Hussain et al. (2021b). "Impact of the COVID-19 outbreak on the US equity sectors: Evidence from quantile return spillovers". In: *Financial Innovation* 7.1, pp. 1–23.

Shannon, Claude E (1948). "A mathematical theory of communication". In: *The Bell system technical journal* 27.3, pp. 379–423.

Sheldon, George, Martin Maurer, et al. (1998). "Interbank lending and systemic risk: An empirical analysis for Switzerland". In: *Swiss Journal of Economics and Statistics (SJES)* 134, pp. 685–704.

Siegel, Sidney (1956). "Nonparametric statistics for the behavioral sciences." In:

Smailović, Jasmina et al. (2013). "Predictive sentiment analysis of tweets: A stock market application". In: *International Workshop on Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*. Springer, pp. 77–88.

Song, Yan-Yan and LU Ying (2015). "Decision tree methods: applications for classification and prediction". In: *Shanghai archives of psychiatry* 27.2, p. 130.

Souza, Thársis TP and Tomaso Aste (2019). "Predicting future stock market structure by combining social and financial network information". In: *Physica A: Statistical Mechanics and its Applications* 535, p. 122343.

Souza, Thársis Tuani Pinto et al. (2015). "Twitter sentiment analysis applied to finance: A case study in the retail industry". In: *arXiv preprint arXiv:1507.00784*.

Steinbacher, Matjaz, Mitja Steinbacher, and Matej Steinbacher (2013). *Credit contagion in financial markets: A network-based approach*.

Stone, Philip J, Dexter C Dunphy, and Marshall S Smith (1966). "The general inquirer: A computer approach to content analysis." In:

Tang, Huifeng, Songbo Tan, and Xueqi Cheng (2009). "A survey on sentiment detection of reviews". In: *Expert Systems with Applications* 36.7, pp. 10760–10773.

Tetlock, Paul C (2007). "Giving Content to Investor Sentiment: The Role of Media in the Stock Market". In: *The Journal of Finance* 62.3, pp. 1139–1168.

Tetlock, Paul C, Maytal Saar-Tsechansky, and Sofus Macskassy (2008). "More than Words: Quantifying Language to Measure Firms' Fundamentals". In: *The Journal of Finance* 63.3, pp. 1437–1467.

Tirea, Monica and Viorel Negru (2013). "Investment portfolio optimization based on risk and trust management". In: *2013 IEEE 11th International Symposium on Intelligent Systems and Informatics (SISY)*. IEEE, pp. 369–374.

Upper, Christian and Andreas Worms (2004). "Estimating bilateral exposures in the German interbank market: Is there a danger of contagion?" In: *European Economic Review* 48.4, pp. 827–849.

Uysal, Alper Kursat and Serkan Gunal (2014). "The impact of preprocessing on text classification". In: *Information Processing & Management* 50.1, pp. 104–112.

Valle-Cruz, David et al. (2020). *Does Twitter Affect Stock Market Decisions? Financial Sentiment Analysis in Pandemic Seasons: A Comparative Study of H1N1 and COVID-19*.

Wan, Xingchen et al. (2021). "Sentiment correlation in financial news networks and associated market movements". In: *Scientific reports* 11.1, pp. 1–12.

Watts, Duncan J and Steven H Strogatz (1998). "Collective dynamics of 'small-world'networks". In: *nature* 393.6684, pp. 440–442.

West, David (2000). "Neural network credit scoring models". In: *Computers & operations research* 27.11-12, pp. 1131–1152.

Wu, Xindong et al. (2013). "Data mining with big data". In: *IEEE transactions on knowledge and data engineering* 26.1, pp. 97–107.

Yin, Hui, Shuiqiao Yang, and Jianxin Li (2020). *Detecting topic and sentiment dynamics due to COVID-19 pandemic using social media*. arXiv preprint arXiv:2007.02304.

You, Quanzeng and Jiebo Luo (2013). "Towards social imagematics: sentiment analysis in social multimedia". In: *Proceedings of the thirteenth international workshop on multimedia data mining*, pp. 1–8.

Zhang, Dayong, Min Hu, and Qiang Ji (2020). "Financial markets under the global pandemic of COVID-19". In: *Finance Research Letters*, p. 101528.

Zhang, Qingchen et al. (2018). "A survey on deep learning for big data". In: *Information Fusion* 42, pp. 146–157.

Zhang, Xue, Hauke Fuehres, and Peter A Gloor (2011). "Predicting stock market indicators through twitter "I hope it is not as bad as I fear"". In: *Procedia-Social and Behavioral Sciences* 26, pp. 55–62.

Zheludev, Ilya, Robert Smith, and Tomaso Aste (2014). "When can social media lead financial markets?" In: *Scientific Reports* 4, p. 4213.