University of Pavia

Doctoral Thesis

Six papers on computational methods for the analysis of structured and unstructured data in the economic domain

Author:
Giancarlo Nicola

Supervisor:
Dr. Paola Cerchiello

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

in the

DREAMT program
Department of Economics and Management

May 20th, 2019

Declaration of Authorship

I, Giancarlo NICOLA, declare that this thesis titled, “Six papers on computational methods for the analysis of structured and unstructured data in the economic domain” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- I have made clear where the thesis is based on work done by myself jointly with others.

Signed: ____

Date: ____

UNIVERSITY OF PAVIA

Abstract

Economics
Department of Economics and Management

Doctor of Philosophy

Six papers on computational methods for the analysis of structured and unstructured data in the economic domain

by Giancarlo NICOLA

This work investigates the application of computational methods for structured and unstructured data. The domains of application are two closely connected fields with the common goal of promoting the stability of the financial system: systemic risk and bank supervision.

The work explores different families of models and applies them to different tasks: graphical Gaussian network models to address bank interconnectivity, topic models to monitor bank news and deep learning for text classification. New applications and variants of these models are investigated posing a particular attention on the combined use of textual and structured data. In the penultimate chapter is introduced a sentiment polarity classification tool in Italian, based on deep learning, to simplify future researches relying on sentiment analysis.

The different models have proven useful for leveraging numerical (structured) and textual (unstructured) data. Graphical Gaussian Models and Topic models have been adopted for inspection and descriptive tasks while deep learning has been applied more for predictive (classification) problems. Overall, the integration of textual (unstructured) and numerical (structured) information has proven useful for systemic risk and bank supervision related analysis. The integration of textual data with numerical data in fact, has brought either to higher predictive performances or enhanced capability of explaining phenomena and correlating them to other events.

UNIVERSITY OF PAVIA

Executive Summary

Economics
Department of Economics and Management

Doctor of Philosophy

Six papers on computational methods for the analysis of structured and unstructured data in the economic domain

by Giancarlo NICOLA

This thesis work investigates the application of computational methods for structured and unstructured data to the economic domain. Due to the abundance of unstructured data generated, there is great interest in their combination with traditional data sources, normally structured data. Unstructured data, differently from structured data, are not organized via pre-defined data models or schema, therefore, specific algorithms are needed to extract their internal structure and organize them. Once organized, the data can be leveraged in combination with traditional structured data to solve a particular task. Specifically, the methods investigated in this work are network models for timeseries analysis and text analytics models (topic models and deep learning). The domains of application are two closely connected fields with the common goal of promoting the stability of the financial system: systemic risk and bank supervision.

The work delves into methodologies that leverage structured and unstructured textual data both separately and combined. It explores different families of models and applies them to tackle different tasks: graphical Gaussian network models to address bank interconnectivity, topic models to monitor bank news and deep learning for text classification. New applications and variants of these models are investigated posing a particular attention on the combined use of textual and structured data. Finally, the thesis aims also to ease future researches leveraging Italian sentiment analysis. In the penultimate chapter is introduced a python package for sentiment polarity classification in Italian based on deep learning. The computational methods investigated in the thesis can also be adapted to different contexts allowing the generalizability of this research towards other task where the integration of textual data can be useful. The work is articulated in nine chapters in which six papers on the use of structured and unstructured textual data for systemic risk and bank supervision are presented.

The first chapter presents an introduction to the problem and to the works exposed in the rest of the manuscript. It follows, in Chapter 2, a literature review of the relevant concepts, methodologies and models explored in the subsequent chapters. The literature review is divided in five sections on Systemic Risk, Contagion, Gaussian Graphical Models, Topic models and Deep Learning for Natural Language Processing. In each section the theory and the relevant literature regarding the topic is reviewed in order to provide the reader with a theoretical support.

In Chapter 3 is presented a framework for systemic risk estimation, based on Graphical Gaussian Models. Two different data sources are incorporated in the stochastic network model, financial markets data and financial tweets, suggesting a way to combine them with a Bayesian approach. The result is a systemic risk estimation model that has been applied to the Italian banking system incorporating network effects and taking into account both twitter and market data.

In Chapter 4 is tested the causal effect between measures derived from a graphical model fitted on US banks stock prices and several financial stress indexes. The graphical model is fitted on market data with a very fast recently developed algorithm based on filtering networks. This allows to update the graphical model for each single market day. Granger causality is then applied to test the causality between the measures computed from the graphical model and the financial stress indexes. The methodology has been applied to the U.S. banking system allowing to calculate a corresponding network model and identifying several system and bank level measures correlating and Granger-causing financial stress indexes.

Chapter 5 presents a work focusing on financial news spreading. It explores how the topics covered in financial news evolve over time among the different considered countries. A causal effect in the diffusion of news is investigated by means of a Granger causality test among topic proportion time-series. The application of a Structured Topic Model, taking into account numerical and categorical covariates as well, allows to incorporate a geographical dimension into the analysis. The Granger causality analysis has evidenced several causal relations in the news diffusion among the different countries.

In Chapter 6 the existence of a causal link between Italian banks' stocks and sentiment analysis on the same bank tweets is investigated. The causality is tested by means of a Granger causality test between the timeseries of twitter sentiment polarity scores and the price, volume, volatility senior and subordinated CDS spreads and subordinated bond spreads of the stocks. The results show that both Twitter sentiment and Twitter volume do significantly affect several financial variables for some of the banks in the sample in particular, those that have recently experienced many episodes of high volatility and negative news.

Chapter 7 exposes a model for bank supervision that combines financial variables and textual news data through a deep learning architecture. In this work the combination of news data and financial structured data is leveraged to improve bank distress predictions. The model has shown to be able to learn the combinations of banks' financial conditions and news semantic content more frequently associated with distress conditions. This is reflected in the improved performance obtained when leveraging both news data and financial numerical data as input to the model.

Chapter 8 introduces a sentiment analysis tool for Italian based on deep learning and discusses the details of the model, its training process and its evaluation. The results of the participation to the ABSITA 2018 challenge from the EVALITA conference along with an analysis on its structure and how it processes the textual input are presented. The chapter includes also a brief guide to the installation and use of the python package implementing the model which is available for download.

Chapter 9 finally presents the conclusions of the thesis discussing the principal lessons learned from the analysis and the possible impact of ongoing researches on the combination of structured data and unstructured textual data.

In conclusion, all of these models, applied to solve different problems have proven to be a valid choice for leveraging numerical (structured) and textual (unstructured) data. Graphical Gaussian Models and Topic models have been successfully adopted for inspection and descriptive tasks while deep learning has been applied more for predictive (classification) problems. Throughout the different works presented, the integration of textual (unstructured) and numerical (structured) information has proven useful for systemic risk and bank supervision related analysis. Depending on the task in fact, the integration of textual and structured data has brought either to higher predictive performances or enhanced capability of explaining phenomena and correlating them to other events.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Structured and unstructured data

Nowadays, social media, mobile applications, smart homes, Internet of Things, financial transactions and other technologies are generating an unprecedented amount of multistructured data. Thanks to falling sensor prices, fast internet connections, large availability of computational power and storage space, an increasing amount of data is generated, recorded and stored every day. These key differential factors, compared to some decades ago, have triggered an exponential growth of data. These data contain useful information that can be analyzed to describe and make predictions regarding the involved phenomena and can be used to improve organizations, systems and products. The analysis to convert them in actionable insights and to distil useful models involve different procedures depending on their structure and characteristics. Algorithms are quite specific on the type and structure of information they work on, thus, different types of data are treated with different methods.

There are two main families in which data can be categorized, structured and unstructured. Structured data are information organized in a given structure and comprised of clearly defined data types whose patterns makes them easily searchable both with human generated queries and via algorithms using type of data and field names. Unstructured data instead, while retaining some internal structure, are not structured via pre-defined data models or schema and therefore they are not as easily searchable. A characteristic of the current pervasive origination of data from several endpoints is the abundance of unstructured data. Around 80% of the data generated today are unstructured and they comprehend texts, images, sensors stream data, videos and audio files, each one of these requiring different pre-processings and algorithms.

This thesis work investigates computational methods for structured and unstructured data applied to the economic domain. More specifically, the methods explored are network models for timeseries analysis and text analytics models (topic models and deep learning). The financial domains of application are two closely connected fields with the common goal of promoting the stability of the financial system: systemic risk and bank supervision. Systemic risk is the risk of severe instability or collapse of the financial system triggered by a single institution failure. It captures the risk of a cascading failure caused by interlinkages between the financial institutions and it has been a major contributor to the financial crisis of 2008. The financial crisis has stressed the necessity of understanding the financial system as a network of institutes, where cross-bank linkages play a fundamental role in the spread of systemic risks. Financial network models, that take into account these complex interrelationships, seem to be an appropriate tool in this context where the focus is posed on the risk contagion and spillover between banks. These models allow to understand which banks are the most interconnected and how they affect each other. Moreover, it is possible to integrate network models derived from different data sources in a model that aggregates different views on systemic risk. Banking supervision has the goal to ensure that banks are operating safely and soundly and following all the rules and regulations. This is done identifying the risks in all areas of operations of the banks (credit risk, liquidity risk, operational risk, capital risk, interest rate risk, profitability risk, reputational risk, internal controls, corporate governance, anti-money laundering) in a timely fashion, and ensuring the stability of credit institutions and of the system through effective action. Both systemic risk estimation and bank supervision can take advantage of the information contained in the news and in the sentiment of financial operators. Text analytics models can be used to process text coming from news and social media to convert it in structured data actionable for systemic risk estimation and bank supervision. In particular, textual data can be leveraged as an additional source of information to complement traditional ones in the analysis. The information contained in texts in fact, are often complementary to those reflected in standard financial databases (i.e. financial ratios, sector indicators, macroeconomic data). While the last, in general, contain very accurate and objective but lower frequency information, text is characterized by higher frequency, higher noise and a certain degree of subjectivity. Considering these differences it is interesting to combine them taking advantage of both their aspects. However, text in the form of natural language, has many peculiarities and requires appropriate techniques for its processing and its inclusion in the analysis. Natural language, intended as the tool that people use to express themselves, is inherently difficult to interpret and "quantify". In fact, it has specific properties that reduce the efficacy of textual information compared to classic structured data, like linguistic variation and ambiguity. By linguistic variation it's meant the possibility of using different words or expressions to communicate the same concept. Linguistic ambiguity instead is related to the several nuances that words or a phrases can have, allowing for different interpretations. Natural Language Processing (NLP) is the field at the intersection of Computer Science, Linguistics and Machine Learning that is concerned with enabling computers to interpret and generate human natural language. In this work we will resort to Natural Language Processing for extracting the information contained in text and then combine them with different data sources.

1.2 Motivation and objective of this thesis

In this dissertation we explore ways to leverage structured data and unstructured textual data for systemic risk estimation and bank supervision. The motivation is to provide methodologies for the combination of structured and unstructured textual data with the objective of benefiting from it in terms of phenomena explicability and predictability. In doing so we consider different class of models. We resort to graphical Gaussian network models when addressing bank interconnectivity, to topic models when monitoring bank news and deep learning when classifying text. We investigate new applications and variants of these models posing a particular attention on the combined use of textual and structured data. In the last chapter is also presented a tool for sentiment analysis in Italian. The aim here is to ease future researches leveraging Italian sentiment analysis. Furthermore, the generalizability of this research towards other tasks and domains where textual data can be exploited, combined or not with structured data, plays an important role. In fact, both the methodologies and the computational tools investigated are of general purpose and can be adapted to different contexts.

1.3 Structure

The manuscript is divided in nine chapters. The first is represented by this introduction to the themes considered and to the work exposed in the rest of the manuscript. It follows, in Chapter 2, a literature review of the relevant concepts, methodologies and models explored in the subsequent chapters. The literature review is divided in five sections on Systemic

Risk, Contagion, Gaussian Graphical Models, Topic models and Deep Learning for Natural Language Processing.

In Chapter 3 is presented a framework for systemic risk estimation, based on Graphical Gaussian Models. Two different data sources are incorporated in the stochastic network model, financial markets data and financial tweets, suggesting a way to combine them with a Bayesian approach.

In Chapter 4 is tested the causal effect between measures extracted from a graphical model fitted on US banks stock prices and several financial stress indexes. The graphical model is fitted with a very fast recently developed algorithm based on filtering networks. This allows to update the graphical model for each single market day. Granger causality is then applied to test the causality between the measures extracted from the graphical model and the financial stress indexes.

Chapter 5 presents a work focusing on financial news spreading. It explores how the topics covered in financial news evolve with time among the different countries. A causal effect in the diffusion of news is investigated by means of a Granger causality test among topic time-series. The application of a Structured Topic Model, taking into account numerical and categorical covariates as well, allows to incorporate a geographical dimension into the analysis.

In Chapter 6 the existence of a causal link between Italian bank stocks and twitter sentiment analysis on the same banks is investigated. The causality is tested by means of a Granger causality test between the timeseries of twitter sentiment polarity scores and the price, volume and volatility of the stocks.

Chapter 7 exposes a work on bank supervision that combines financial variables and textual news data through a deep learning model. In this work the combination of news data and financial structured data is leveraged to improve bank distress predictions.

Chapter 8 introduces a sentiment analysis tool for Italian based on deep learning. The details of the model, training process and evaluation are discussed. The results of the participation to the ABSITA 2018 challenge from the EVALITA conference along with an analysis on its structure and how it processes the textual input are presented. The chapter includes also a brief guide to the installation and use of the python package implementing the model which is available for download.

Chapter 9 finally presents the conclusions of this work discussing the principal lessons learned from the analysis and the possible impact of ongoing researches on the combination of structured data and unstructured textual data.

1.4 Publications

Work contributing to this thesis has been published in the following peer reviewed Journals and conferences:

- Chapter 3: Cerchiello, P., Giudici, P. and Nicola, G. (2017). Twitter data models for bank risk contagion, In Neurocomputing, Volume 264, Pages 50-56, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2016.10.101.

- Chapter 5: Cerchiello, P., Nicola, G., (2018). Assessing News Contagion in Finance. Econometrics 6 (1), pages 5–24. 10.3390/econometrics6010005.

- Chapter 8: Nicola, G. (2018). Bidirectional Attentional LSTM for Aspect Based Sentiment Analysis on Italian. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18), Turin, Italy. CEUR.org. - Best Single-Author Student Contribution at the EVALITA'18 conference.

1.5 Collaborations

Work contributing to these Chapters has been fruit of the following collaborations:

- Chapter 3: Paola Cerchiello and Paolo Giudici from University of Pavia.
- Chapter 4: Tomaso Aste from University College London (UCL) and Paola Cerchiello from University of Pavia.
- Chapter 5: Paola Cerchiello from University of Pavia
- Chapter 6: Giuseppe Bruno and Juri Marcucci from Bank of Italy and Paola Cerchiello from University of Pavia.
- Chapter 7: Paola Cerchiello from University of Pavia, Samuel Rönnqvist from University of Turku and Peter Sarlin from Hanken School of Economics.

Chapter 2

Literature review

This chapter is meant to provide a common background to the works presented in the upcoming chapters. It gives a brief introduction to the concepts and the literature of Systemic Risk and Contagion and revises the theory and the literature of the methodologies and models applied in this work. It is divided in five sections on Systemic Risk, Contagion, Gaussian Graphical Models (GGM), Topic models and Deep learning. The introductory sections on Systemic Risk and Contagion revise their definitions and the different interpretations available in literature. The subsequent sections, related to statistical models, expose the theory on which the models are based on, the relevant literature and the connection with their application in the following chapters.

2.1 Systemic Risk

The recent financial and economic crisis has made evident the critical role played by the financial system interconnections in channeling and amplifying shocks. In fact, a distinguishing trait of the last crisis has been the impact of systemic risk and of systemic effects arising from the existing linkages among banks and financial institutions. Broadly speaking, systemic risk refers to the risk that financial instability becomes so widespread that it impairs the functioning of a financial system to the point where economic growth and welfare suffer materially [ECB 2009]. While it's difficult to give a precise definition of systemic risk and currently there is not an entirely commonly accepted definition of systemic risk, it is possible to recall some of its more widely adopted descriptions. Systemic risk can be described as the risk of experiencing a strong systemic event that adversely affects several systemically important intermediaries or markets (including potentially related infrastructures). The event could be triggered by both endogenous shocks from within the financial system or the economy at large or exogenous shocks external to the financial system. The exogenous shock in turn can be systematic (widespread) or idiosyncratic (limited in scope). The resulting systemic event, in turn, can be considered weak or strong if it involves failure of concerned intermediaries or dysfunctionality of the affected markets. Moreover, it is possible to differentiate between a "horizontal" view of systemic risk, where the attention is confined to the financial system, and a "vertical" view of systemic risk where the two-sided interactions between the financial system and the economy at large are considered [De Bandt and Hartmann 2000]. When it comes to assessing the severity of systemic risk and systemic events one possible way is to look at the effects that they have on the real economy in terms of consumption, investments and growth or economic welfare.

Systemic risk is a complex phenomenon due to its interaction effects, and the several distinctions between idiosyncratic or systematic factors, exogenous or endogenous triggers and sequential or simultaneous impacts illustrate the complexity of this phenomenon. To restrict and delimit the possible research directions resulting from the combination of these elements normally three main "forms" of systemic risk are considered: the contagion risk, the risk of macroeconomic shocks inducing simultaneous problems and the risk of the unravelling of imbalances in the system. These three main forms of systemic risk which are not mutually exclusive can realize independently or jointly. Contagion usually refers to a supposedly idiosyncratic problem that becomes more widespread in the cross-sectional dimension, often in a sequential fashion. A contagion example is the failure of a bank triggering the failure of other banks that initially seemed solvent. The second type of systemic risk refers to the risk of widespread exogenous shocks simultaneously affecting several intermediaries or markets. Materializations of this risk are the bank crisis that have been witnessed in conjunction with macroeconomic shocks like cyclical downturns, interest rate hikes and stock market crashes [Gorton 1988, Lindgren et al. 1996]. The third type of systemic risk stems from the unravelling of widespread imbalances built-up endogenously over time within the financial system which may adversely affect several intermediaries or markets simultaneously.

2.2 Contagion

In this Section we revise the concept and definition of contagion, in view of its importance for the topics covered in this work, as a major driver and justification for bank regulation. We discuss its definition from the viewpoint of the application domain of systemic risk and bank supervision and its connection to related terms like spillover and interdependence. Most of the literature on economic contagion deals with contagion among countries, banks, within large-value payment systems and among major financial markets. [Forbes and Rigobon 2002] define contagion as a shock transmission between markets (or portion of markets) where there was no previous dependence prior to the shock. Thus, contagion is regarded as a marked increase in the market dependence following to a shock event. In fact, even if two markets are highly linked one to another and share a high degree of correlation but this hasn't changed after the shock, the phenomenon can't be considered contagion but rather integration. [Bekaert et al. 2005] describe contagion as an excess of correlation between markets, more than it can be explained by economic fundamentals. Both these definitions come with the necessity to identify the normal degree of dependence during periods of stability, together with the fundamentals. [Corsetti et al. 2001] consider contagion as a break in the parameters governing the correlation system. According to them, in fact, it's normal to have some comovement across markets following a shock caused by global or regional factors, and such phenomena is not contagion but can be regarded as a consequence of interdependence. For example, a rise in volatility of asset prices in one market can be expected to correlate with a rise of volatility in other markets due to international interconnectedness and the related transmission mechanisms. However, when contagion takes place, this degree of transmission is very high and exceeding what can be predicted whit a constant transmission mechanism, and it's propagated mainly by irrational investor behaviours.

A slightly different approach is taken in [Kaminsky et al. 2003] where contagion is regarded as an instantaneous effect following a shock propagating rapidly between the markets. According to this definition, the fundamental characteristic is the speed of diffusion of financial distress. When the propagation to the other markets is gradual, then the event can't be regarded as contagion, but rather as a spillover episode. The term spillover also indicates the phenomenon in which an event in one context is triggered from other events in seemingly unrelated contexts. Anyway, compared to contagion it lacks the "unexpected" or "surprising" component of the transmission of shocks across markets. In the category of spillovers are included the effects of common shocks throughout all the markets, such as changes in reference interest rates and energy commodity prices. Thus, spillover effects are transmissions of financial distress due to interdependence among markets. Economic interdependence is a phenomenon that stems from having multiple economic agents depending on each other to source the necessary inputs to produce their outputs. It's characteristic of societies with a high degree of division of labor where the agents cannot produce all the goods that they need to realize their products. The interdependence can be at different level, individuals, companies and countries. From the 1950s, global economic interdependence has grown exponentially, as a result of a great technological progress and policies aimed at opening national economies internally and externally to global competition that finally resulted in increased international trading flows. When the actors of an economic system need to participate to a trading network to obtain the products they cannot produce efficiently for themselves, economic interdependence is generated among the network participants. The high degree of interdependence of modern global economy has been a key driver in the evolution and the spreading of the last financial and economic crisis.

Returning to contagion, a different approach to its definition, in contrast with the previous, relates it directly to investor expectations and behaviours. [Masson 1998] for example, associates pure contagion to changes in investors' expectations unrelated to the macroeconomic fundamentals of a country, often identified as monsoonal effects. [Karolyi and Stulz 2006] in disagreement with this view, instead relate contagion to investors irrational and panic behaviour when a shock is propagated from one market to another. They consider that market contagion can be defined regardless of whether it is transmitted through macroeconomic fundamentals or not. Another point of view is given in [Engle 2009] where the contagion channels, next to the fundamentals, are found within the investor behaviours, and traced to the portfolios that they trade within multiple markets. This theory grounds on the fact that volatilities and the correlations between asset returns and stock markets depend on the information available to the investors. In fact, assets are hold by investors in anticipation of the payment that are to be made in the future, so the asset value is fundamentally linked to the forecast on its future price evolution and the news regarding the market are what make investors change their future price forecast, as formulated in the model of changing asset prices by [**?**]. So, countries with similar economies are correlated because they are influenced by the same events and the same news will drive investors to decisions. According to this approach, contagion can be defined as a sudden shift in investor's market expectations or confidence.

As we have seen there are many definitions of contagion in literature and all of them frame this phenomenon from different points of view. Even within the same author literature is possible to witness gradual shifts over time in the definition of contagion, like in [Rigobon 2016], where the author in contrast with [Forbes and Rigobon 2002] uses the words "contagion" and "spillovers" as describing very loosely the phenomenon in which a shock from one country is transmitted to another. Despite the little convergence about the definition of contagion there are two main characteristics that seem to be prominent and they are the suddenness and unexpectedness of the shock transmission in the contagion process. When referring to this contagion in the rest of this work we refer to it considering these two main properties.

2.3 Gaussian Graphical Models

Graphical models are a general framework to compactly represent large joint probability distributions using a set of 'local' relationships among neighbouring variables in a graph [Darroch et al. 1980, Kindermann et al. 1980, Lauritzen 1996, Jordan et al. 1999]. Several commonly used statistical models (e.g. Kalman filters, Hidden Markov Models, Ising Models) can be described as graphical models. In fact, they have been successfully applied to a broad range of problems, from computer vision to financial modelling to computational biology. Graphical models provide a principled approach for dealing with uncertainty through the use of probability theory, and an effective approach for coping with complexity through the use of graph theory. In these models, the graph structure defines the conditional independence properties of the underlying probability distributions. It encodes the information about which variables influence each other. This allows to answer questions like: are variables X and Y dependent because they "interact" directly, or because they are both dependent on a third variable Z? The advantage of Graphical models consist in efficiently representing a joint distribution P over some set of random variables $X = \{X_1,...,X_n\}$. In fact, even in simple cases where these variables are binary-valued, a joint distribution requires the specification of 2^n numbers (the probabilities of the 2^n different assignments of values $x_1,...,x_n$). However, it is often the case that there is some structure in the distribution that allows to factor its representation into modular components. Graphical models precisely provide a general-purpose modelling language for exploiting the independence properties of distributions, which exist also in many real-world phenomena. The independence properties in fact, can be leveraged to represent such high-dimensional distributions more compactly and this compact parametrization is what enables the model learning. Finally, inference in Graphical models provides the mechanisms for gluing all these components back together in a probabilistically coherent manner. The two most common subclasses of graphical models are the Bayesian networks (also known as belief networks or causal networks) and Markov networks (also known as Markov random fields (MRFs)). The representation of Bayesian networks is based on directed graphs and hence they are also called directed graphical models. The representation of Markov networks instead, is based on undirected graphs and hence they are called also undirected graphical models. Mixed directed and undirected representation (see, for example, the work on chain graphs [Lauritzen and Wermuth 1989, Buntine 1995]) are possible but less common. The main research problems regarding Graphical models are: representation, sampling, inference and learning. In [Jensen 1996, Bishop 2006, Koller and Friedman 2009] there exhaustive reviews of the several research questions associated with graphical models. Representation concerns choosing the right model for the considered application. Sampling considers the problem of generating samples from the model's joint probability distribution. Inference regards using the model to answer probabilistic queries like, computing marginal probabilities or inferring the value of unobserved variables. Model learning focuses on recovering the model from data, estimating its structure and the values of its parameters. This is a very common task because in many settings it is necessary to estimate a model from a dataset with no details about the model. In this chapter we will focus on Markov networks and in particular, Gaussian graphical models that are the class of models applied in the next chapters.

2.3.1 Markov networks

Markov networks (or MRFs) are based on undirected graphical models. They are useful in modelling a variety of phenomena where is not possible to naturally ascribe a directionality to the variables interaction. Furthermore, the undirected model often proves itself simpler in terms of independence structure and inference task compared to directed models. The nodes in the undirected graph of a Gaussian graphical model represent the variables, and the edges correspond to some notion of direct probabilistic interaction between the neighbouring variables. The graph structure represents the qualitative properties of the distribution. To completely represent the distribution, it is necessary to associate the graph structure with a set of parameters. In Gaussian graphical models, partial correlations can be estimated assuming that the observations follow a multivariate Gaussian, in which the covariance matrix Σ is constrained by the conditional independences described by a graph [Lauritzen 1996].

More formally, let $X = (X_1,...,X_N) \in R^N$ be a $N-$ dimensional random vector distributed according to a multivariate normal distribution $\mathcal{N}(\mu,\Sigma)$. Without loss of generality, we will assume that the data are generated by a stationary process, and, therefore, $\mu = 0$. In addition, we will assume throughout that the covariance matrix Σ is not singular.

Let $G = (V, E)$ be an undirected graph, with vertex set $V = \{1, ..., N\}$, and edge set $E = V \times V$, a binary matrix, with elements e_{ij}, that describe whether pairs of vertices are (symmetrically) linked between each other ($e_{ij} = 1$), or not ($e_{ij} = 0$). If the vertices V of this graph are put in correspondence with the random variables $X_1, ..., X_N$, the edge set E induces conditional independence on X via the so-called Markov properties [Lauritzen 1996]. In particular, the pairwise Markov property determined by G states that, for all $1 \leq i < j \leq N$:

$$e_{ij} = 0 \Longleftrightarrow X_i \perp X_j | X_{V \setminus \{i,j\}}; \tag{2.1}$$

that is, the absence of an edge between vertices i and j is equivalent to independence between the random variables X_i and X_j, conditionally on all other variables $x_{V \setminus \{i,j\}}$.

Let the elements of Σ^{-1}, the inverse of the variance-covariance matrix, be indicated as $\{\sigma^{ij}\}$, [Whittaker 1990] proved that the following equivalence also holds:

$$X_i \perp X_j | X_{V \setminus \{i,j\}} \Longleftrightarrow \rho_{ijV} = 0 \tag{2.2}$$

where

$$\rho_{ijV} = \frac{-\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}} \tag{2.3}$$

denotes the ij-th partial correlation, that is, the correlation between X_i and X_j, conditionally on the remaining variables $X_{V \setminus \{i,j\}}$.

Therefore, by means of the pairwise Markov property, and given an undirected graph $G = (V, E)$, a graphical Gaussian model can be defined as the family of all N-variate normal distributions that satisfies the constraints induced by the graph on the partial correlations, as follows:

$$e_{ij} = 0 \Longleftrightarrow \rho_{ijV} = 0 \tag{2.4}$$

for all $1 \leq i < j \leq N$.

Finding the right model or models underlying the data is equivalent to finding the graph or graphs G which best represent the conditional independences between the different variables. Stochastic inference in graphical models may lead to two different types of learning: structural learning, which implies the estimation of the graphical structure G that best describes the data, and quantitative learning, that aims at estimating the parameters of a graphical model, for a given graph. In a Bayesian framework, structural learning can be achieved choosing the graphical structure on the basis of the posterior distribution for G or the marginal likelihood for the model corresponding to G [Berger 1985]. If we use the [Diaconis and Ylvisaker 1979] conjugate prior distribution for the precision matrix K, the marginal likelihood is actually equal to the ratio of norming constants of two Wishart distributions conditional on having those entries corresponding to the missing edges of G equal to zero. Such Wishart distributions will be called G-Wishart. The norming constant of the G-Wishart with shape parameter δ and inverse scale parameter D is therefore equal to the integral over the set of positive definite matrices with zero ij entries whenever $(i, j) \notin E$, of a Wishart density

$$g(K) \sim det(K)^{\frac{\delta-2}{2}} e^{-\frac{1}{2}tr(KD)} \tag{2.5}$$

This norming constant is also needed for the computation of the Bayes factors in model comparisons and for any Markov chain or stochastic search on the space of all possible graphs on p vertices. Its efficient and accurate computation is therefore very important. While when G is complete or decomposable, this norming constant is well known and can be obtained analytically, when G is non-decomposable, it has to be computed numerically. Considering this if we can now recall the expression of the likelihood of a graphical Gaussian model.

For a given graph G, consider a sample X of size n. For a subset of vertices $A \subset N$, let Σ_A denote the variance-covariance matrix of the variables in X_A, and define with S_A the corresponding observed variance-covariance sub-matrix.

When the graph G is decomposable (and we will assume so) the likelihood of the data, under a graphical Gaussian model, nicely decomposes as follows [Dawid and Lauritzen 1993]:

$$p(X|\Sigma, G) = \frac{\prod_{C \in \mathscr{C}} p(X_C|\Sigma_C)}{\prod_{S \in \mathscr{S}} p(X_S|\Sigma_S)}, \tag{2.6}$$

where $X_{\mathscr{C}}$ and $X_{\mathscr{S}}$ respectively denote the set of random variables belonging to the cliques and to the separators of the graph G, and where:

$$P(X_C|\Sigma_C) \propto |\Sigma_C|^{-n/2} exp[-\frac{1}{2} tr\left(S_C\left(\Sigma_C\right)^{-1}\right) \tag{2.7}$$

and similarly for $P(X_S|\Sigma_S)$.

This is a key result since it allows to calculate analytically the likelihood of a proposed model during the structural learning and will be used in the next chapters for Gaussian graphical model selection.

2.4 Topic Models

Statistical document interpretation stems from the necessity to organize large volumes of documents in an easily searchable structure. For this, several statistical methods have been developed to represent texts in a compact way. Among them, probabilistic topic models, a class of statistical methods for interpreting the contents of document collections, have proven very useful. These models help understanding and organizing corpora by learning sets of topics from words frequently co-occurring in documents.

Probabilistic topic models are the last stream of a research that grounds on semantic representations of documents. Among these representations one of the first and widely adopted is the Vector Space Model (VSM) [Salton 1975] also known as semantic space or Bag-of-Words (BoW) representation, where documents and terms are represented by vectors over an Euclidean space. The most common type of semantic space is the term-document matrix [Salton 1975, Salton and McGill 1983, Turney and Pantel 2010]. The term-document matrix C incorporates information regarding the frequency of terms within the corpus documents. Given a vocabulary of W terms and a collection of D documents it will have a size of $W \times D$. It's elements, c_{ij} are the counts of the occurrences of the i^{th} word in the j^{th} document. Often they are also weighted using the TF-IDF (Term Frequency – Inverse Document Frequency) weighting [Salton and McGill 1983]. However, pure VSMs are characterized by high dimensionality and sparsity of the semantic space. The high dimensionality is caused by large number of unique terms that normally appears in texts and the sparsity is due to the fact that many terms appear only in few documents. These problems impact the accuracy when calculating the similarity between documents or terms.

These limitations can be mitigated by reducing the high dimensionality and sparsity of the term-document matrix with methods like Latent Semantic Analysis (LSA) [Deerwester et al. 1990, Landauer and Dumais 1997]. LSA decompose the term-document matrix C into three other matrices applying a Singular Value Decomposition (SVD):

$$C = U\Sigma V^T \tag{2.8}$$

where U is the word vectors matrix with size $W \times W$ whose columns are the eigenvectors of CC^T, Σ is the diagonal $W \times D$ matrix of the singular values and V is the document vectors matrix of size $D \times D$ whose columns are the eigenvectors of $C^T C$. The multiplication of the three component matrices results in the original matrix C if the number of singular values is no smaller than the smallest dimension of C (i.e. the matrix is perfectly decomposed). However, LSA reduces the dimensionality of the semantic space, by deleting part of the singular values in the diagonal matrix Σ starting from the smallest one. Doing so, the three matrices product becomes just an approximation of the original matrix C. In this way it is possible to reduce the dimensionality of the original semantic space. The resulting approximated matrix obtained by keeping only the K largest singular values is the following:

$$C \approx U_K \Sigma_K V_K^T \tag{2.9}$$

where U_K is the word vectors matrix with size $W \times K$, Σ_K is the $K \times K$ diagonal matrix of the K singular values and V_K is the document vectors matrix of size $K \times D$. [Stevens 2012] shows that LSA can be interpreted already as a topic model. In fact, it learns a set of topics T_K obtained by multiplying the word vectors matrix U_K with the diagonal matrix of the singular values Σ_K:

$$T_K = U_K \Sigma_K \tag{2.10}$$

Moreover, LSA returns also the topic assignments for each document, given by the matrix V_K^T. However, topics learned by LSA are not easily interpretable. In fact, the topic vectors are linear combinations of the term-document frequencies consisting both of positive and negative values [Stevens 2012] preventing a straightforward interpretation as unnormalized probabilities. This limitation is overcome by the probabilistic topic models that produce more descriptive and coherent topics compared to LSA.

2.4.1 Probabilistic Topic Models

Introduced around the early 2000s in [Hofmann 1999, Blei et al. 2003], probabilistic topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents [Blei 2012]. While texts analysis is their main application, topic models algorithms can be adapted to different types of data. Among other applications, they have been used to find patterns in genetic data, images, and social networks.

From a statistical modelling point of view they are generative models that learn a set of latent variables called topics. Topic models are based on the assumption that documents are generated by a mixture of topics and that topics are defined by probability distributions over words. Normally documents are represented as "Bag of Words" (BoW), ignoring word order and taking into account only the terms count per document. The only information that is relevant to topic models is the number of times a word appears in each document. Collections of documents are then represented as term-document matrices where the terms count for each document are recorded. The topic models receive in input a term-document matrix (representative of the corpus) and return as output a set of topics together with topic assignments to the documents. Usually, to calculate the term-document matrix, documents are tokenized (split up) into words or phrases and converted to their lower-case form.

Each topic is represented by a probability distribution over all the unique tokens (words) in the vocabulary. Tokens that co-occur frequently in documents will be assigned a high probability in the same topics and likely to represent a coherent subject. Each document is represented as a probability distribution over topics with only a few topics assigned with high probability. Topic models can be used to organize large text collections by clustering documents under different topics and to infer the topic themes by inspecting their soft-clustered terms.

Probabilistic generative models can be represented using the so-called plate notation, a convenient method for representing variables that repeat in a graphical model (see e.g. Figures 2.1, 2.2, 2.3). In this representation, shaded nodes indicate observed variables while unshaded nodes indicate latent variables. Conditional dependency among variables is represented by arrows while plates (boxes) surrounding nodes indicate repetitions of sampling steps with the number of repetitions indicated in the bottom right corner of the plate.

2.4.2 Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (PLSA) is a topic model presented by [Hofmann 1999] based on a mixture decomposition derived from a latent class model. In PLSA each word in a document, containing N words, is modelled as a sample from a mixture model built on a set of T topics in the form of multinomial random variables. Like in the other topic models structure, words represent the observed variables while topics represent the latent variables. To fit the model, once the number T of topic is fixed, it's necessary to calculate the probability distribution over words for each topic and the probability distribution over topics for each document.

Let assume a topic ϕ_z is a distribution over a fixed vocabulary of size V. In the original PLSA model, the distribution ϕ while not explicitly specified is in the form of a Multinomial distribution. Thus, ϕ_z is a vector that represents the topic distribution over words.

Let also assume that a document consists of multiple topics. Therefore, there is a distribution θ_d over a fixed number of topics T for each document d. This distribution also takes the form a Multinomial distribution, where each element represents the probability that topic z appears in document d.

The topics obtained from the algorithm are the result of the following generative process which is represented in plate notation in Figure 2.1:

1. For each document d within a corpus D with probability $P(\theta, d)$,
2. For each word w_n of document d with $n \in 1,..,N$:
   - select a latent topic z with probability $P(z|d)$,
   - select a word w with probability $P(w|z)$.

The expression of the joint probability that defines the above process is the following:

$$P(\theta_d, w) = P(\theta_d) P(w|\theta_d) \tag{2.11}$$

$$P(w|\,\theta_d) = \sum_{z \in Z} P(w,z) P(z|\theta_d) = \tag{2.12}$$

$$P(\theta_d, w) = P(\theta_d) \sum_{z \in Z} P(w,z) P(z|\theta_d) \tag{2.13}$$

Where, $P(z|\theta_d)$ is the topic distribution of document d and $P(w|z)$ is the word distribution of topic $z \in Z$. given a document d and each document consists of multiple topics. This process, respects the main assumption of topic models by representing each document directly as a list of topic weights. This leads to several problems: (1) the number of parameters in the model grows linearly with the size of the corpus, which leads to serious problems with overfitting, and (2) it is not clear how to assign probability to a document outside of the training set [Blei et al. 2003].

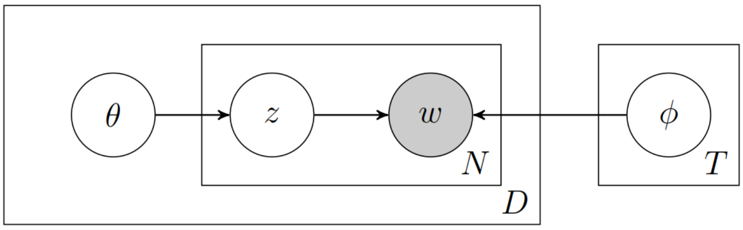


FIGURE 2.1: pLSA model representation in plate notation

2.4.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) has been presented in [Blei et al. 2003] as an extension of PLSA to overcome the two problems mentioned above. To this purpose, LDA introduces symmetric Dirichlet priors on both the topics distribution for a particular document, θ, and on the words distribution for a particular topic, ϕ. In LDA the topic weights for each document are treated as a hidden random variable with T, equal to the number of topics. The topics obtained from the algorithm are the result of the following generative process which is represented in plate notation in Figure 2.2:

1. For each topic t in T
   - Draw a word distribution over words $\phi \sim Dir(\beta)$
2. Choose N, the number of words in a document (depends on the vocabulary)

   A document d within a corpus D is represented by latent topics through the following generative process [Blei et al. 2003]:
3. Draw the topic distribution for a document $\theta \sim Dir(\alpha)$
4. For each word w_n with $n \in 1,..,N$:
   - Choose a topic $z_n \sim Multinomial(\theta)$
   - Choose a word w_n from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n

where:

- N is the number of words in a document.
- z_n is the topic for the word w_n.
- θ is the distribution over topics for a document.
- α is the parameter of the Dirichlet prior on the per-document topic distributions.
- β is the parameter of the Dirichlet prior on the per-topic word distribution.

The following equation describes the joint probability of the corpus D given the corpus level hyperparameters α and β:

$$P(D|\alpha,\beta) = \prod_{t=1}^{T}\prod_{d=1}^{D}\prod_{n=1}^{N} P(\phi_t|\beta)P(\theta_d|\alpha)P(z_{dn}|\theta_d)P(w_{dn}|\phi_{z_{dn}}) \tag{2.14}$$

The per-topic word distributions ϕ and the per-document topic distributions θ are the main variables that need to be estimated in the model. Since Inferring direct estimates from

Equation 2.14 is intractable, different optimization methods have been applied to the problem. [Hofmann 1999] used the expectation-maximisation (EM) algorithm to estimate ϕ and θ directly. However, the EM algorithm may get stuck in local maxima, thus approximation methods like Bayesian Variational Inference [Blei et al. 2003] and Gibbs sampling [Griffiths and Steyvers 2004] have been used to overcome this problem. Gibbs sampling is an algorithm of the Markov Chain Monte Carlo (MCMC) family for sampling values from multivariate probability distributions. Gibbs sampling is commonly used for statistical inference, especially Bayesian inference, as an alternative to deterministic algorithms for statistical inference such as the expectation-maximization algorithm (EM) [Dempster et al. 1977]. Gibbs sampling initially assigns each word w to a random topic $t \in \{1,...,T\}$ for every document in corpus D. Then, for each word, it estimates the probability of allocating that current word in each topic, given the topic allocations of the other words. The calculation for this probability is provided in [Griffiths and Steyvers 2004]:

$$P(z_i = j | z_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha} \tag{2.15}$$

with $z_i = j$ representing the topic allocation of word w_i in topic j, z_{-i} being the topic allocations of the other words, and $\cdot$ representing the remaining information from words, documents and hyperparameters α and β. The element C^{WT}_{wj} of the matrix CVT, with size $V \times T$, counts the number of times that word w has been allocated to topic j. The element C^{DT}_{dj} of the matrix CDT, with size $D \times T$, represents the number of times topic j is assigned to words in document d. The words distribution per topic and the topic distributions per per document are given by the elements of matrices Φ and Θ, which can be obtained by the following equations:

$$\phi_{ij} = \frac{C^{WT}_{ij} + \beta}{\sum_{k=1}^{W} C^{WT}_{kj} + W\beta} \qquad \theta_{jd} = \frac{C^{DT}_{dj} + \alpha}{\sum_{k=1}^{T} C^{DT}_{dk} + T\alpha} \tag{2.16}$$

where ϕ_{ij} represents the probability of word w_i in topic j and θ_{jd} represents the probability of topic j in document d_d.

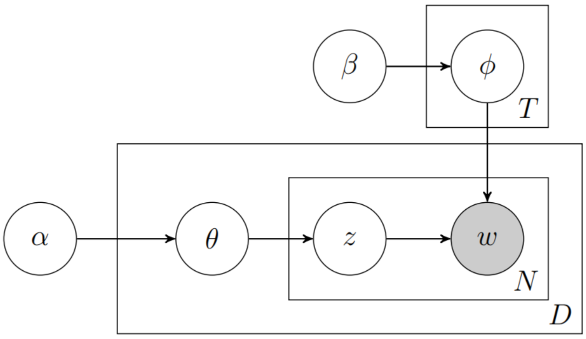


FIGURE 2.2: LDA model representation in plate notation

2.4.4 Correlated Topic Models

The Correlated Topic Model (CTM) [Blei and Lafferty 2006] overcomes one limitation of LDA, the independence hypothesis among topics, which leads to ignore possible inter-topic correlations. The Correlated Topic Model, models these correlations through a logistic normal distribution over topics. Let $\{\mu, \Sigma\}$ be a K-dimensional mean and covariance matrix, and let topics $\phi_{1:T}$ be T multinomials over a fixed word vocabulary of size W. The Correlated

Topic Model, whose plate notation representation is shown in Figure 2.3, assumes that an N-word document arises from the following generative process [Blei and Lafferty 2006]:

1. Draw $\theta|\{\mu,\Sigma\} \sim \mathcal{N}(\mu,\Sigma)$
2. For each word w_n with $n \in 1,..,N$:
   - Choose a topic $z_n|\eta$ from $Multinomial(f(\theta))$
   - Choose word $w_n|\{z_n, \phi 1 : T\}$ from $Multinomial(f(\phi_{zn}))$

The only difference between this process and the generative process of LDA is that the topic proportions are drawn from a logistic normal rather than a Dirichlet. The Correlated Topic Model is more expressive than LDA because removes the strong independence assumption imposed by the Dirichlet in LDA. This hypothesis is not realistic when analyzing document collections, where it's possible to find strong correlations among topics [Blei and Lafferty 2006]. The covariance matrix of the logistic normal distribution defines a topic graph where each node represents a topic and the edges represent the correlations among them.

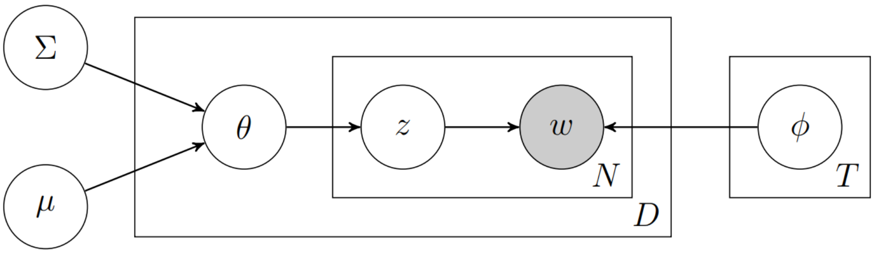


FIGURE 2.3: CTM representation in plate notation

2.5 Deep Learning for Natural Language Processing

Deep Learning is a class of machine learning models that employ multiple processing layers to learn hierarchical representations of data. While many classic machine learning methods require an extensive feature engineering to make accurate predictions, deep learning tries to jointly learn good data representations (features), across multiple level of increasing complexity and abstraction, and the final prediction [LeCun et al. 2015]. The automated feature learning allows an easier automation of the entire learning process, reducing the need for time consuming feature handcrafting and simplifying the model application to different tasks. Deep learning draws its origins from the early research on Neural Networks (NNs) [McCulloch and Pitts 1943, Rosenblatt 1958, Ivakhnenko 1967]. After that, many important milestones have contributed to evolve the early neural networks in what we call today deep learning models. Some of these milestones have been the adoption of the backpropagation algorithm for the training [Rumelhart et al. 1986], the identification of the fundamental deep learning problem (vanishing-exploding gradients) [Hochreiter 1991], the introduction of the Long Short Term Memory (LSTM) networks for long sequences, the introduction of the convolutional neural networks for images and the normalized initialization strategy for the layer weights [Glorot and Bengio 2010]. All these theoretical contributions (and many more, the list is not exhaustive) gradually have made neural networks training easier and faster. This, along with a great availability of computational power (through CPUs, GPUs and TPUs) and open source automatic differentiation frameworks (e.g. Pytorch and Tensorflow) has promoted the application of deep learning and its growing diffusion since the early 2000's.

In this Section we introduce the basics concepts of neural networks with a particular attention to the architectures used in NLP.

2.5.1 Feed-forward Neural Network

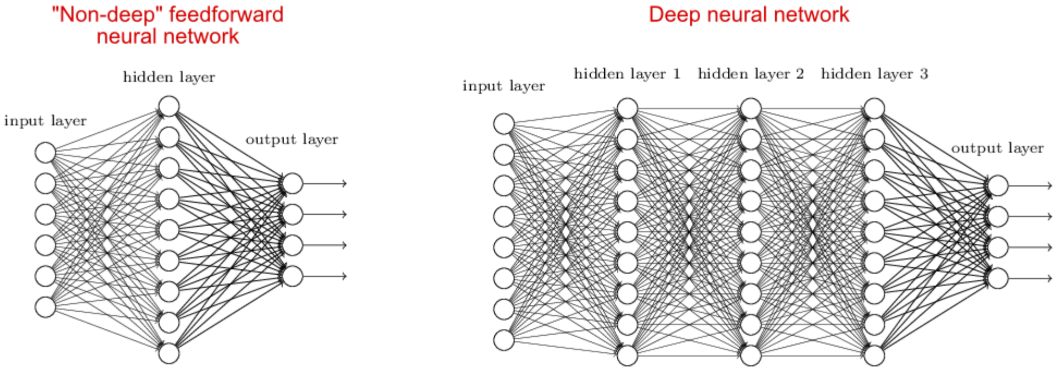


FIGURE 2.4: Schematic representations of a shallow neural network (left) and of a deep neural network (right) [Nielsen 2015]

Neural networks are composed by a sequence of layers (represented vertically in Figure 2.4), each layer containing one or more units (or neurons, represented with circles in Figure 2.4). In the network in Figure 2.4, each unit is connected to every unit the following layer (vice-versa a unit/neuron is connected to every unit/neuron in the previous layer). The connections are represented by lines, that define some weighting of the output of the unit at the start of the line (from left), for the input of the unit at the end of the line. Each layer can be described mathematically as a vector of activations. In the case of the first layer (the input layer, on the left side of both the networks represented in Figure 2.4), these activations are equal to the input. In the final layer (output layer, on the right side of both the networks represented in Figure 2.4) which encodes the output of the network, we would like these activations to be such to minimize the error of the model. The collection of all connections (lines) between two layers, can be mathematically described as a matrix of weights (e.g. W_0). The application of the weight matrix to the input of the network correspond to a matrix multiplication and can be described in linear algebraic notation as

$$\vec{z}_1 = W_0\vec{x} \tag{2.17}$$

where $\vec{z}_1$ is the vector resulting from this linear transformation. The $\vec{z}$-vectors resulting from the matrix multiplication are called pre-activation vectors. Each hidden layer thus, first encodes the sum of the multiplications of each of the activations in the previous layer by some weight in the pre-activation vectors. Then, to obtain the output of each unit (neuron) in the layer, a non linear activation function is applied to the pre-activation vector,

$$\vec{a}_1 = f(\vec{z}_1) \tag{2.18}$$

where f is the non-linear activation function. Figure 2.5 summarizes the entire calculation that take place in a single unit, which consists of inputs x_i, bias unit b, an activation function f and the output.

The output of the unit is computed by the following function:

$$\vec{a} = f(W^T x + b) \tag{2.19}$$

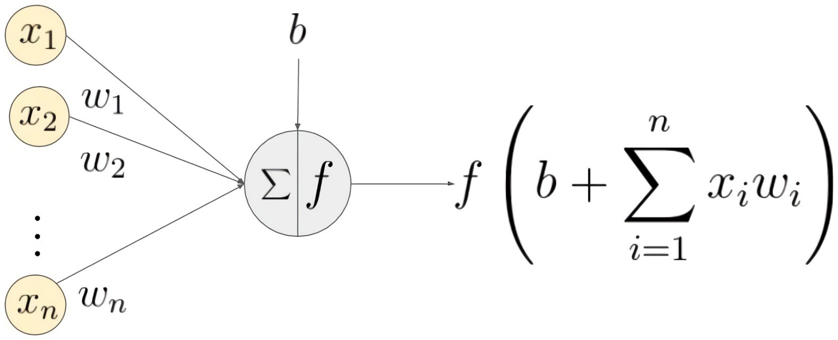


FIGURE 2.5: Schematic representation of the calculations that take place in a neural network neuron. The scheme shows the input $(x_1,...,x_n)$, their corresponding weights$(w_1,...,w_n)$, a bias b and the activation function f applied to the weighted sum of the inputs [Medium]

where f is the activation function, also called a nonlinearity. The first commonly used activation function was the sigmoid function:

$$f(x) = sigmoid(x) = \frac{1}{1+e^{-x}} \tag{2.20}$$

A neural network stacks several of such single units (also called neurons) to compose a layer (vertically in Figure 2.4) and stacks multiple layers in sequence followed by a final output layer. For multiple units each neuron's activation a_i is computed by one inner product with its parameters, followed by an addition with its bias: $a_i = f(W_i \cdot x + b_i)$, where each parameter vector $W_i \in R^n$. It is possible to disentangle the multiplication and the nonlinearity and write this in matrix notation for m many units stacked in a layer as:

$$z = wx + b \quad \text{and} \quad a = f(z) \tag{2.21}$$

where $W \in R^{m \times n}$, $b \in R^m$ and the function f is applied elementwise:

$$f(z) = f([z_1, z_2, ..., z_m]) = [f(z_1), f(z_2), ..., f(z_m)] \tag{2.22}$$

The output of such a neural network layer can be seen as a transformation of the input that captures various interactions of the original inputs. Thus, Neural networks can be seen as a collection of non-linear functions applied to a series of matrix-vector multiplications, mapping from one domain (e.g., words) to another (e.g., sentence polarity). This most basic neural network variant, depicted in Figure 2.4 is called Feed-forward Neural Network (FFNN) (also known as a multilayer perceptron) and was first introduced in 1958 in [Rosenblatt 1958]. The architectural difference between a simple FFNN and a deep learning FFNN it's only in the number of layers the network is composed of.

2.5.2 Activation functions

Each hidden unit applies an activation function to the sum of its weighted inputs, as written in equation 2.19. Different functions can be used as activation functions but they should have two main properties:

1. The function needs to be non-linear, since it is for this kind of functions it has been proven that a feed-forward network with a single hidden layer containing a finite number of neurons is a universal function approximator [Cybenko 1989].
2. The function should be monotonic since the error surface associated with a single-layer model will then be convex [Wu 2009].

Some of the most commonly used activation functions that meets these requirements are listed in Table 2.1.

Name	Function
Logistic (Sigmoid)	$f(x) = \frac{1}{1+e^{-x}}$
Hyperbolic tangent (tanh)	$f(x) = \frac{e^{2x}-1}{e^{2x}+1}$
Rectified Linear Unit (ReLU)	$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x, & \text{for } x \geq 0 \end{cases}$
Leaky ReLU	$f(x) = \begin{cases} 0.01x & \text{for } x < 0 \\ x, & \text{for } x \geq 0 \end{cases}$
Softmax	$f(x) = \frac{e^{x_i}}{\sum_{k=1}^{K} e^{x_k}}$ for $i = 1, ..., K$

TABLE 2.1: List of activation functions most commonly used in neural networks

Other activations are often preferred in practice over the traditionally popular logistic function due to better empirical performance like computational speed and stability to the vanishing gradient problem. In certain cases activation functions are inspired by biological neurons like the Rectified Linear Unit (ReLU). The nonlinearity applied by the ReLU it's in part similar to what happens in a biological neuron [Hahnloser et al. 2000, Hahnloser and Seung 2001]. That is to say, when the input is below a certain threshold, the neuron does not fire, and when the input is above this threshold, the neuron fires with a current proportional to the input. ReLUs are nowadays commonly used in many neural network architectures since they suffer less from the vanishing gradient problem and they have been found to make it substantially easier to train deep networks [Nair and Hinton 2010]. The main disadvantage of ReLUs is that they can end up in state in which they are inactive for almost all inputs, meaning that no gradients flow backward through the unit. This is known as "dying ReLU" problem and can be mitigated by using leaky ReLUs activation functions, for which even negative input have some negative activity associated, preventing complete inactivity of the unit and allowing for error propagation. A particular case is the softmax function that in general is used as the final layer activation for classification problems. In fact, softmax has the property of normalizing the activation on an entire layer to 1 yielding a probability distribution based on its input. Moreover, by construction it makes easy for the network to assign much more probability to one class than to the the others thus, it is commonly used in classification problems were the classes are mutually exclusive.

2.5.3 Training

During the training phase the weights of the neural network are adjusted in order to minimize the loss function calculated on the training examples. Practically due to computational constraints the only optimization methods used nowadays for neural networks are gradient based. Thus, the training of the model consists of an iterative procedure in which each iteration is composed by two sequential phases, first a “forward pass” and then a “backward pass”. The training phase in its complex is a continuous succession of forward and backward passes. During the forward pass the input is fed into the network and the algebraic operations and nonlinearities map the input to the prediction to calculate the value of the loss function. Then in the backward pass the error signal (value of loss function) is backpropagated into the layers of the network to adjust the weights. The two passes have opposite directions, in the forward pass the information and the operations flow from the input layer to the output layer while in the backward pass they flow from the output layer back to the input layer. The weights adjustment takes place in the backward pass where each trainable weight of the network is modified by a quantity proportional to the gradient of the error with respect to that weight. The key actors of this process are the Loss function that gives the error signal, the Stochastic Gradient Descent (SGD) (or one of its recent variants) that defines the magnitude of the weight adjustment (given the gradient of the error with respect to that weight) and the backpropagation algorithm that "backpropagates" the gradients from the output layer of the network back to the inner layers. In the next sections we well briefly introduce these arguments.

2.5.4 Loss function

Deep learning models are trained adjusting the network weights on the base of an error signal provided by a loss function. Different loss functions can be defined to measure the model error for different type of problems. In general, the loss functions applied in NNs fall into two classes depending upon the prediction task: “classification losses” applied in classification problems (i.e. when attempting to predict some discrete class label, out of a finite set of labels), and “regression losses” applied in regression problems (i.e. when attempting to predict some continuous score). These two settings account for most of deep learning applications and classification particularly is the most common case in NLP (e.g. in POS tagging, NER, language identification, machine translation). In classification problems, the activation function of the output layer of the network is the softmax function that allows to interpret the network prediction as a probability distribution over the different classes. In these tasks normally, the loss function applied to calculate the error is the cross-entropy between the output of the softmax activations and the target labels probability distribution:

$$L(\vec{\hat{y}},\vec{y}) = -\sum_{i=1}^{N} \vec{y}_i \log \vec{\hat{y}}_i \tag{2.23}$$

where L denotes the loss function, $\vec{y}$ is the target probability distribution over labels, $\vec{\hat{y}}$ is the model’s predicted distribution given an input x. The more different the two distributions, the highest the error of the model. In regression problems, instead the most commonly applied loss function is the Mean Squared Error (MSE), defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.24}$$

where $\hat{y}$ is the predicted label, and y is the true label. This function is commonly used in regression, and is especially handy for explaining backpropagation, as in the next section. While these losses can be applied to most of the problems, there are many more complicate loss functions that can be applied to specific cases and, in general, defining the loss function to apply is one of the most important and impactful step in training a model.

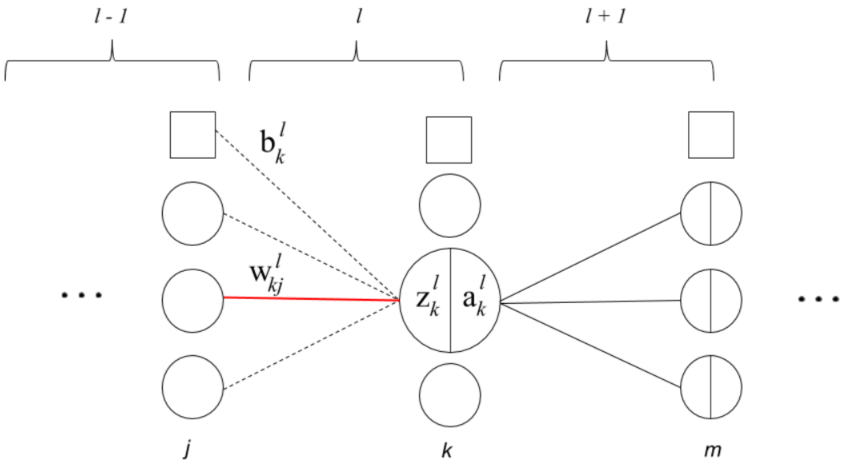


FIGURE 2.6: Schematic representation of a block of the three layers in a neural network for the illustration of the backpropagation algorithm. Three layers anywhere in the network, derivative is taken with respect to the weight shown in red. The middle neuron is enlarged for visualization purposes [Medium]

2.5.5 Backpropagation algorithm

The Backpropagation algorithm [Bryson et al. 1963, Werbos 1974, Rumelhart et al. 1986] allows to calculate the gradient of the loss function for each weight which can be then used by the optimization algorithm (like SGD) to update the weight. It is referred to also as Backpropagation of errors because the error is calculated at the output and distributed back through the network layers, following the inverse direction of the forward pass. Training the network is about understanding how changing the weights and biases in a network changes the cost function, which consists in computing the partial derivatives $\partial C/\partial w$ and $\partial C/\partial b$ of the cost function C with respect to any weight w or bias b in the network. Recalling the equations to calculate a single unit activation:

$$z = wx + b \quad \text{and} \quad a = f(z) \tag{2.25}$$

and the following network scheme:

The input sum of a neuron k in layer l is expressed as:

$$z_k^l = \sum_j w_{kj}^l a_j^{l-1} + b_k^l \tag{2.26}$$

and similarly, the input sum of a neuron m in layer $l+1$ is expressed as:

$$z_m^{l+1} = \sum_k w_{mk}^{l+1} a_k^l + b_m^{l+1} \tag{2.27}$$

Then, the derivative of the error function with respect to a weight connecting a unit j to unit k in layer l can be expressed using the derivatives chain rule as:

$$\frac{\partial C}{\partial w_{kj}^l} = \frac{\partial C}{\partial z_k^l}\frac{\partial z_k^l}{\partial w_{kj}^l} = \frac{\partial C}{\partial a_k^l}\frac{\partial a_k^l}{\partial z_k^l}\frac{\partial z_k^l}{\partial w_{kj}^l} \tag{2.28}$$

$$= \left(\sum_m \frac{\partial C}{\partial z_m^{l+1}}\frac{\partial z_m^{l+1}}{\partial a_k^l}\right)\frac{\partial a_k^l}{\partial z_k^l}\frac{\partial z_k^l}{\partial w_{kj}^l} = \tag{2.29}$$

$$= \left(\sum_m \frac{\partial C}{\partial z_m^{l+1}} w_{mk}^{l+1} \right) f' \left(z_k^l \right) a_j^{l-1} \tag{2.30}$$

To compute these derivatives, we first introduce an intermediate quantity, δ_{lk}, which we call the output error in the k^{th} unit. Backpropagation gives us a procedure to compute the error δ_{lk} and relate δ_{lk} to $\partial C / \partial w_{kj}^l$.

$$\delta_k^l \equiv \frac{\partial C}{\partial z_k^l} \tag{2.31}$$

Substituting this definition in Equation 2.30 we have a recursive formula for the error signals, that allows us to backpropagate the error:

$$\delta_k^l = \left(\sum_m \delta_m^{l+1} w_{mk}^{l+1} \right) f' \left(z_k^l \right) \tag{2.32}$$

Similarly, for the bias the error function can be derived with respect to them as well, remembering that $b_k^l = 1$:

$$\frac{\partial C}{\partial b_k^l} = \frac{\partial C}{\partial z_k^l} \frac{\partial z_k^l}{\partial b_k^l} = \delta_k^l \tag{2.33}$$

We can see that the gradient of the cost function with respect to the bias for each unit is simply the unit's error signal. To apply this recursive formula, we need the first error signal to backpropagate that is the error signal of the neurons in the output layer L of the network.

$$\delta_j^L = \frac{\partial C}{\partial z_j^L} = \frac{\partial C}{\partial a_j^L} \partial a_j^L \partial z_j^L = \frac{\partial C}{\partial a_j^L} f' \left(z_j^L \right) \tag{2.34}$$

Having this we can "propagate" backwards through the network calculating all the error signals to update the weights according to the applied optimization algorithm. From a computation point of view, it is interesting to note that derivatives computed for higher layers are reused when computing derivatives for lower layers, making the process very efficient. Moreover, if it is created a library of differentiable functions or layers where for each function is known how to forward-propagate (directly applying the function) and how to back-propagate (knowing the function's derivative), it is possible to compose any complex neural network. It is only necessary to keep a stack of the function calls during the forward pass and their parameters in order to know the way back to backpropagate the errors using the derivatives of these functions. This can be done by de-stacking through the function calls. This technique is called auto-differentiation, it requires only that each function is provided with the implementation of its derivative and is at the basis of deep learning frameworks like Theano, Tensorflow and Pytorch.

2.5.6 Stochastic Gradient Descent

The most common optimization algorithms for neural networks training are the Stochastic Gradient Descent (SGD) and its variants (Adagrad, RMSProp, Adam, and others [Hinton et al. 2012, Kingma and Ba 2014]. In SGD, the gradient for weight updates is calculated based on a minibatch of n samples drawn from the training set [Bottou 1998], differently form the gradient descent algorithm that approximates the gradient with the entire training set. This is done for memory constraints when the training set is too large to fit entirely in the optimization device memory. As a result, SGD updates are noisier compared to gradient descent updates and they point only on average the parameter updates in the direction of maximum decrease of the loss function. The updates for a weight are defined by the following equation:

$$w := w - \lambda \nabla C(w) = w - \lambda \frac{1}{n} \sum_{i=1}^{n} \nabla C_i(w) \tag{2.35}$$

Where w is the weight to optimize, λ is the learning rate, C is the cost function and n is the minibatch size. Interesting variants of the SGD are Adagrad and RMSProp that adjust the learning rate for each weight and the Adam algorithm that is similar to RMSProp and has shown better results compared to the others on many problems [Kingma and Ba 2014]. While the goal of the optimization algorithm is to tune the model parameters to find the global minimum of the cost function, with deep networks there is no guarantee of achieving this. Most of the times during the training the network ends up in a local minimum of the cost function. However, in the case of supervised learning with deep neural networks, most local minima appear to have a low loss function value, roughly equivalent to that of the true global minimum [Saxe et al. 2013, Dauphin et al. 2014, Goodfellow et al. 2015, Choromanska et al. 2015].

2.5.7 Network initialization

Parameter initialization is fundamental for neural networks stability during training. One of the key factors that have greatly improved NNs trainability in the last ten years are the strategies developed for weights initialization. NNs must be initialized with different weights among the units of the same layer otherwise a zero-like initialization would result in all hidden units representing the same function and, due to how backpropagation works, receiving the exact same weight updates. Therefore, it is necessary to initialize the weights according to a distribution with non-zero variance. The work of [Glorot and Bengio 2010] has introduced the idea of rescaling the variance of the distribution according to the number of hidden units in the layers that the weights connect. In fact, their strategy is to initialize each weight with a small Gaussian value with zero-mean and variance based on the fan-in and fan-out of the weight. Other commonly used methods include those introduced by and [Saxe et al. 2013], by [He et al. 2015] when using the ReLU activation function and by [Goodfellow et al. 2016] often used in the case of recurrent neural networks for which particular care needs to be taken.

2.5.8 Regularization in Neural Networks

Regularization is an important tool to contrast overfitting during neural network training. Overfitting prevents the model from generalizing well to new examples and it's more likely when the training examples are scarce or when the network is heavily overparametrized for the task. There are several methods to reduce overfitting like L^1 or L^2 regularization but the most used nowadays in deep networks is the dropout mechanism [Srivastava et al. 2014]. The key idea of dropout is to randomly drop units (along with their connections) with a predetermined probability from the neural network during training. The same units are dropped during both the forward and backward pass. This prevents units from co-adapting too much, forcing the network to select more robust features and to be less reliant on specific units. Dropout can be regarded as a form of ensembling similar to bagging [Breiman 1994], in fact during training, it samples from an exponential number of different "thinned" networks. At test time, it is easy to approximate the effect of averaging the predictions of all these thinned networks by simply using a single unthinned network that has proportionally smaller weights. For recurrent neural networks, like for the initialization, special dropout variants

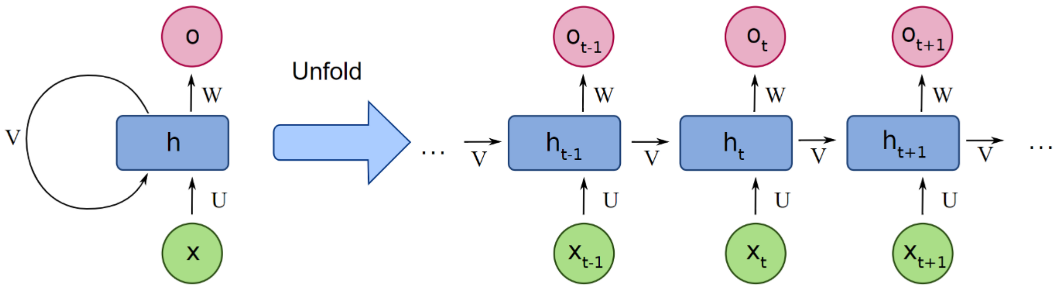


FIGURE 2.7: Representation of a Recurrent Neural Network in its folded and unfolded form [Deloche 2017]

can be applied like recurrent dropout [Semeniuta et al. 2016], or variational dropout [Gal and Ghahramani 2016] in which the same dropout mask is used for each time step.

2.5.9 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are deep learning models designed for dealing with sequential data [Elman 1990]. They are very deep feedforward neural network that have a hidden layer receiving its own input for each timestep and share weights (U, V in Figure 2.7) across time steps.

They are dynamic models that map sequences to sequences working on sequential inputs of arbitrary length $(x_1,x_2,...,x_n)$ and outputting another sequence $(y_1,y_2,...,y_n)$. With a sufficient number of hidden units, an RNN can approximate any measurable sequence-to-sequence mapping to arbitrary accuracy [Hammer 2000] (this is the equivalent of the universal approximation theorem for FFNN). Each o_t in the output sequence can take advantage of the information in the input sequence up to step t thanks to the recurrent connections. In fact, they allow a "memory" of previous inputs to persist in the network's internal state, which can then be used to influence the network output. On the left side of Figure 2.7 the RNN is represented as an FFNN with a loop, whereas on the right side it's depicted in its unrolled (unfolded) version. The output of RNNs is given by the following equations:

$$\vec{z}_t = W_s \vec{x}_t, \tag{2.36}$$

$$\vec{s}_t = f_s(\vec{z}_t), \tag{2.37}$$

$$\vec{y}_t = f_y(W_y \vec{s}_t) \tag{2.38}$$

where W_s is the matrix of weights for the current time step's input (xt), U is the weight matrix for the connections from the previous time step, st is the internal state vector of the network representing the history of the sequence, t is the index of the current time step, Wy is the matrix of weights for the output, and fs and fy are the activation functions applied to the preactivations of internal state and to the preactivations of the output. This architecture is referred to as Elman net, or Simple RNN [Elman 1990]. The advantage of having the internal state s, is that the network can leverage preceding information when calculating the output $\vec{y}_t$. In practice, the prediction at each time step is conditioned on the inputs of the entire preceding sequence. RNNs however, are difficult to train on problems with long-range temporal dependencies [Bengio et al. 1994, Hochreiter and Schmidhuber 1997, Martens and Sutskever 2011] due to their nonlinear iterative nature. In fact, a small change to an iterative process can compound, resulting in very large effects many iterations later. The implication of this is that the derivative of the loss function at one timestep can be exponentially large with respect to the hidden activations at a much earlier timestep making the gradients "explode"

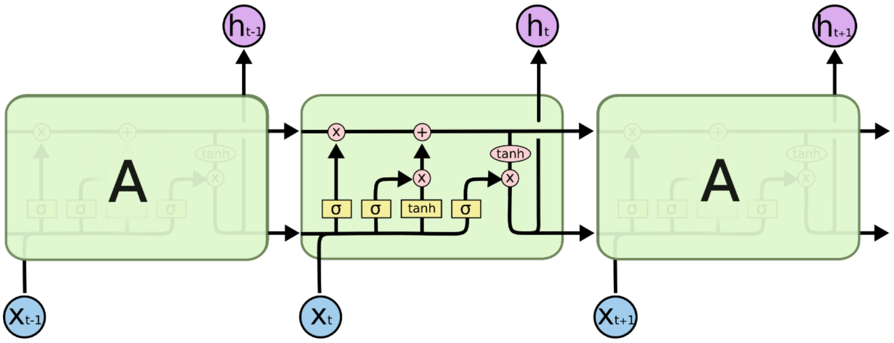


FIGURE 2.8: Schematic representation of a LSTM and of the computation that performs [Olah 2015]

during training; this is known as the exploding gradient problem. Moreover, RNNs also suffer from the opposite problem of "vanishing gradient", first described by [Hochreiter 1991] and [Bengio et al. 1994]. Since the gradients in early layers of the network are the result of many multiplication operators on numbers smaller than one, they become very small and don't provide the training signal. Both these problems derive from the fact that RNN are very deep neural networks and thus suffer from having unstable gradients, as the gradients calculated by backpropagation are dependent on the output of the network, which can be quite far away from the first layers in the network. The vanishing and the exploding gradient problems make it difficult to optimize RNNs on sequences with long-range temporal dependencies.

2.5.10 Long Short Term Memory Networks

Long Short Term Memory (LSTM) networks, introduced in [Hochreiter and Schmidhuber 1997], are an improvement of RNN with memory cells specifically designed to mitigate the problem of unstable gradients. They suffer less from the vanishing and exploding gradient problems and are able to capture long-range dependencies [Cho 2015]. A LSTM network is formed exactly like a simple RNN, except that the nonlinear units in the hidden layer are replaced by memory blocks. Each memory block contains one or more self-connected memory cells and three multiplicative units, the input, output and forget gates, that modify the extent to which old information is remembered or forgotten.

The key idea behind LSTMs is the introduction of components along which information can flow mostly unchanged and be preserved for many time steps. This idea is implemented by the cell state, the horizontal line running through the top of the diagram in Figure 2.8. The cell state in fact, has few linear interactions with the input, output and forget gates that regulate the addition or removal of information to the cell state. The gates are composed by a simple sigmoid neural net layer and a pointwise multiplication operation and are a way to optionally let information through. They allow the memory cells to store and access information over long sequences of time steps. When the gating units are shut (i.e. low activation), the gradients can flow through the memory unit without alteration for an indefinite amount of time, thus overcoming the vanishing gradients problem. A low activation of the input gate (gate closed), prevents new inputs from overwriting the cell activation, that can therefore be read from the net much later in the sequence, by opening the output gate. The behaviour of the gates and of the other vectors of the LSTM are governed by the following equations (for clarity of notation due to the presence of the forget gate and output gate the activation function is indicated with σ instead of f):

$$f_t = \sigma(W_f x_t + U_f y_{t-1} + b_f) \tag{2.39}$$

$$i_t = \sigma(W_i x_t + U_i y_{t-1} + b_i) \tag{2.40}$$

$$o_t = \sigma(W_o x_t + U_o y_{t-1} + b_o) \tag{2.41}$$

$$c_t = f_t c_{t-1} + i_t \sigma_c(W_c x_t + U_c y_{t-1} + b_c) \tag{2.42}$$

$$\hat{y}_t = o_t * \sigma(c_t) \tag{2.43}$$

where f_t represents the output of the forget gate, it represents the output of the input gate, ot represents the output of the output gate, ct represents the cell state, $\hat{y}_t$ represents the output vector, W and U represent the weight matrices, xt represents the current input, and b represents bias units. These are the equations that describe how the input and output gates preserve the contents of the memory unit and how the forget gate can make the memory unit forget its contents.

LSTM have been successfully applied to many real world problems requiring long range memory like reinforcement learning [Bakker 2002], music generation [Eck and Schmidhuber 2002], speech recognition [Graves and Schmidhuber 2005, Graves et al. 2006], protein secondary structure identification [Chen and Chaudhari 2005, Hochreiter et al. 2007], handwriting recognition [Liwicki et al. 2007, Graves et al. 2008], machine translation [Sutskever et al. 2014], robotic control [Mayer et al. 2006], and to solve Partially-Observable Markov Decision Processes [Wierstra and Schmidhuber 2007, Dung et al. 2008].

2.5.11 Bidirectionality

For many sequence labelling tasks, it could be useful to condition the output on both past and future context. For example, when working with language, since many of its properties depend on both preceding and proceeding contexts, it is desirable to condition the network output on both contexts simultaneously. This can be achieved using Bi-directional RNNs. The idea behind Bi-directional RNNs is to feed each training sequence forwards and backwards to two different recurrent hidden layers, which are both connected to the same output layer. This provides the network with a symmetric, past and future context for every timestep in the input sequence [Schuster and Paliwal 1997, Graves and Schmidhuber 2005, Goldberg 2015]. Bi-directional LSTMs have been successfully applied to several NLP tasks, like POS tagging, named entity tagging, and chunking [Wang et al. 2015, Yang et al. 2016, Plank et al. 2016] improving performances compared to standard LSTMs.

2.5.12 Word vector representations

Distributional word representation methods exploit word co-occurrences to build dense vector encodings of words. They are based on the distributional hypothesis which states that words used and occurring in the same contexts tend to purport similar meanings [Harris 1954]. According to this hypothesis each word can be represented by means of its neighbours [Firth 1957]. This idea has proven very useful in NLP and distributional representations (also called word vector representations or word embeddings) are at the basis of most deep learning NLP models used today and also many Bag of words methods ground on it, like LSA [Deerwester et al. 1990] and LDA [Blei et al. 2003]. In these representations each word is associated to a real-valued vector, often tens or hundreds of dimensions. This is contrasted with sparse vector representations, such as a one-hot encoding, that have a single "1" at the index location of the current word and require thousands or millions of dimensions (the size of the vocabulary). Thus, the main advantage of word vector representations over sparse vector representations is that they offer a more expressive and efficient representation by maintaining the context similarity of words and by building low dimensional vectors. Popular distributional analysis methods such as Word2Vec [Mikolov et al. 2013] and

GloVe [Pennignton et al. 2014] have been critical to the success of many recent natural language processing applications [Collobert and Weston 2008, Turney and Pantel 2010, Turian et al. 2010, Socher et al. 2013, Goldberg 2015]. The idea at the basis of such algorithms is to construct a neural network that outputs high scores for windows of words that truly occur in a large unlabelled corpus and low scores for corrupted windows (where for example one word is replaced by a random word). When such a network is optimized via gradient descent the derivatives backpropagate into a word embedding matrix $L \in R^n \times V$, where V is the size of the vocabulary and n the dimensionality of the vector representation. In this way, the word vectors are trained to capture distributional semantics and co-occurrence statistics. The resulting word vector space encodes grammatical and semantic properties (i.e. gender and verb tense) in specific distance vectors, allowing to perform meaningful algebraic operations on words (i.e. adding a "gender changing" vector to another word vector to find its male/female equivalent). The vectors can be learned both in an unsupervised way, relieving the network from the burden of words semantic through the labelled dataset (normally smaller), and in a supervised way, in general for fine-tuning them on specific tasks.

2.6 Summary

In this chapter was revised the foundational literature review of graphical models, topic models and deep learning for NLP in order to provide a theoretical support for the methodologies applied in the next chapters. These methods allow to handle also unstructured data like text and some of them are fit for combining them taking advantage of both. While this is a sufficient introduction to understand the algorithms applied in this thesis, it is by no means a complete account of these topics. For a more in-depth review, I refer the readers to [Lauritzen 1996] and [Koller and Friedman 2009] for graphical models, to [Blei and Lafferty 2009] for topic models and to [Goodfellow et al. 2016] and [Goldberg 2015] for neural networks and their application to NLP.

Chapter 3

Twitter data models for bank risk contagion

3.1 Summary

A very important and timely area of research in finance is systemic risk modelling which concerns the estimation of relationships among different financial institutions. Understanding these relationships can help to establish which institutions are more contagious/subject to contagion. The aim of this chapter is to develop a systemic risk model that includes not only the information from financial market prices, but also the information contained financial tweets. From a methodological viewpoint, we propose a new framework, based on Graphical Gaussian Models, to estimate systemic risks from two different sources and suggest a way to combine them, using a Bayesian approach. From an applied viewpoint, we present a systemic risk model based on financial markets prices timeseries and financial tweets sentiment timeseries and show that such data can shed further light on the interrelationships between financial institutions.

3.2 Introduction

Systemic risk models address the issue of interdependence between financial institutions and, specifically, measure how bank default risks are transmitted among banks.

The study of bank defaults is important for two reasons. First, an understanding of the factors related to bank failures enables regulatory authorities to supervise banks more efficiently. If supervisors can detect problems early enough, regulatory actions can be taken, to prevent a bank from failing and reduce the costs of its bail-in or bail-out. Differently these costs would be mainly faced by shareholders, bondholders and depositors in case of bail-in or governments and, ultimately, taxpayers in case of bail-out. Second, the failure of a bank can induce failures of other banks or of part of the financial system. Understanding the determinants of a single bank failure may thus help to understand the determinants of financial systemic risks, were they due to microeconomic idiosyncratic factors or to macroeconomic imbalances. When problems are detected, their causes can be removed or isolated, to limit "contagion effects".

Most research papers on bank failures are based on financial market models, that originate from the seminal paper [Merton 1974], in which the market value of bank assets is matched against bank liabilities. Due to its practical limitations, Merton's model has been evolved into a reduced form (see e.g. [Vasicek 1984]), leading to a widespread diffusion of the resulting approach, and the related implementation in regulatory models.

The last few years have witnessed an increasing research literature on systemic risk, with the aim of identifying the most contagious institutions and their transmission channels. A comprehensive review is provided in [Brunnermeier and Oehmke 2012]. Specific measures of systemic risk have been proposed for the banking sector; in particular by [Adrian and Brunnermeier 2009], [Acharya et al. 2010](MES), [Brownlees and Engle 2011, Huang et al. 2011], [Acharya et al. 2012](SRISK), [Cao 2013](δCoVaR), [Banulescu and Dumitrescu 2015](CES), [Calabrese and Giudici 2015, Segoviano and Goodhart 2009]. On the basis of market prices, these authors calculate the quantiles of the estimated loss probability distribution of a bank, conditional on the occurrence of an extreme event in the financial market. The above approach is useful to establish policy thresholds aimed, in particular, at identifying the most systemic institutions. However, it is a bivariate approach, which allows to calculate the risk of an institution conditional on another (or on a reference market), but it does not address the issue of how risks are transmitted between different institutions in a multivariate framework.

Trying to address the multivariate nature of systemic risk, researchers have proposed a network modelling approach, following the idea in [Diamond and Dybvig 1983] and the seminal papers of [Sheldon and Maurer 1998, Eisenberg and Noe 2001, Boss et al. 2004, Upper and Worms 2004]. In this literature, interconnectedness is related to the detection of the most central players in a network that describes financial flows between agents. The simplest way of measuring the centrality of a node in the network is by counting the number of its neighbours. However, more sophisticate measures of centrality have been developed, including that shown in [Battiston et al. 2012] who develops a network algorithm, the DebtRank, starting from Google's PageRank algorithm.

A different type of network models, recently proposed, are based on correlations (or distances) between financial descriptors of agents, such as their stock market prices, bond interest rate spreads or corporate default spreads. The first contributions in this framework are [Mantegna 1999, Onnela et al. 2004] and, more recently, [Billio et al. 2012] and [Diebold and Yilmaz 2014], who propose measures of connectedness based on Granger-causality tests and variance decompositions. [Barigozzi et al. 2013, Ahelegbey et al. 2015] and [Giudici and Spelta 2016] have extended the approach introducing stochastic graphical models. Here we shall follow this latter approach, considering a stochastic framework, based on graphical models. We will thus be able to derive, on the basis of market price data on a number of financial institutions, the network model that best describes their interrelationships and, therefore, explains how systemic risk is transmitted among them. It is well known that market prices are formed in 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 systemic risk estimation. This weakness of the market suggests to enrich financial market data with data coming from other, complementary, sources. Indeed, market prices are only one of the evaluations that are carried out on financial institutions. Other relevant ones include ratings issued by rating agencies, reports of qualified financial analysts, and opinions of influential media.

Most of the previous sources are private, and not available to the general public. However, summary reports from these sources are now typically reported, almost in real time, in social networks and in tweets. In parallel with these developments, seminal papers on the statistical analysis of such type of data have recently appeared: see, for example, [Bollen et al. 2011, Bordino et al. 2012, Choi and Varian 2012, Feldman 2013, Cerchiello and Giudici 2015] who all show the added value of tweets and, more generally, of textual data in economics and finance. Indeed, twitter data offer the opportunity to extract complementary information to market prices and that can, in addition, "replace" market information when not available (as it occurs for banks that are not listed). In order to extract these information from tweets it is necessary to preprocess their text with Natural Language Processing techniques. In our context, we rely on sentiment analysis to obtain a daily sentiment score from financial tweets collected daily on a number of Italian banks. These scores are then aggregated in timeseries that express the average daily Twitter users sentiment towards each considered bank.

In this paper we propose to build Graphical Gaussian Models using daily variation of banks "sentiment", and to integrate them with graphical models based on market data, by means of a Bayesian approach. This allows to obtain a comprehensive measurement framework of bank interconnectedness, that can be employed to understand contagion effects. The novelty of this work is twofold. From a methodological viewpoint, we propose a new framework, based on Graphical Gaussian Models, to estimate systemic risks from two different sources and suggest a way to combine them, using a Bayesian approach. From an applied viewpoint, we present a systemic risk model based on financial markets prices timeseries and tweets sentiment timeseries and show that such a model can help updating the default probability of a bank, conditionally on the others.

The rest of the Chapter is organised as follows: in Section 3.3 we introduce our proposal; in Section 3.4 we explore the analyzed financial and tweet data; in Section 3.5 we apply our method to the Italian banking system; finally, in Section 3.6 we present some concluding remarks.

3.3 Methodology

We first introduce the graphical network models that will be used to estimate relationships between banks, both with market and tweet data. Relationships between banks can be measured by their partial correlation, that expresses the direct influence of a bank on another. Partial correlations can be estimated assuming that the observations follow a graphical Gaussian model, in which Σ is constrained by the conditional independences described by a graph (see e.g. [Lauritzen 1996]).

More formally, let $X = (X_1,...,X_N) \in R^N$ be a $N-$dimensional random vector distributed according to a multivariate normal distribution $\mathcal{N}(\mu,\Sigma)$. Without loss of generality, we will assume that the data are generated by a stationary process, and, therefore, $\mu = 0$. In addition, we will assume throughout that the covariance matrix Σ is not singular. Let $G = (V,E)$ be an undirected graph, with vertex set $V = \{1,...,N\}$, and edge set $E = V \times V$, a binary matrix, with elements e_{ij}, that describe whether pairs of vertices are (symmetrically) linked between each other ($e_{ij} = 1$), or not ($e_{ij} = 0$). If the vertices V of this graph are put in correspondence with the random variables $X_1,...,X_N$, the edge set E induces conditional independence on X via the so-called Markov properties (see e.g. [Lauritzen 1996]).

In particular, the pairwise Markov property determined by G states that, for all $1 \leq i < j \leq N$:

$$e_{ij} = 0 \Longleftrightarrow X_i \perp X_j | X_{V \setminus \{i,j\}}; \tag{3.1}$$

that is, the absence of an edge between vertices i and j is equivalent to independence between the random variables X_i and X_j, conditionally on all other variables $x_{V \setminus \{i,j\}}$. Let the elements of Σ^{-1}, the inverse of the variance-covariance matrix, be indicated as $\{\sigma^{ij}\}$, [Whittaker 1990] proved that the following equivalence also holds:

$$X_i \perp X_j | X_{V \setminus \{i,j\}} \Longleftrightarrow \rho_{ijV} = 0 \tag{3.2}$$

where

$$\rho_{ijV} = \frac{-\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}} \tag{3.3}$$

denotes the ij-th partial correlation, that is, the correlation between X_i and X_j, conditionally on the remaining variables $X_{V \setminus \{i,j\}}$.

Therefore, by means of the pairwise Markov property, and given an undirected graph $G = (V, E)$, a Graphical Gaussian Model can be defined as the family of all N-variate normal distributions that satisfies the constraints induced by the graph on the partial correlations, as follows:

$$e_{ij} = 0 \Longleftrightarrow \rho_{ijV} = 0 \tag{3.4}$$

for all $1 \leq i < j \leq N$. Stochastic inference in graphical models may lead to two different types of learning: structural learning, which implies the estimation of the graphical structure G that best describes the data, and quantitative learning, that aims at estimating the parameters of a graphical model, for a given graph. Structural learning can be achieved choosing the graphical structure with maximal likelihood. To this aim, we now recall the expression of the likelihood of a Graphical Gaussian Model. For a given graph G, consider a sample X of size n. For a subset of vertices $A \subset N$, let Σ_A denote the variance-covariance matrix of the variables in X_A, and define with S_A the corresponding observed variance-covariance sub-matrix.

When the graph G is decomposable (and we will assume so) the likelihood of the data, under a graphical Gaussian model, nicely decomposes as follows [Dawid and Lauritzen 1993]:

$$p(X|\Sigma, G) = \frac{\prod_{C \in \mathscr{C}} p(X_C|\Sigma_C)}{\prod_{S \in \mathscr{S}} p(X_S|\Sigma_S)}, \tag{3.5}$$

where $X_{\mathscr{C}}$ and $X_{\mathscr{S}}$ respectively denote the set of random variables belonging to the cliques and to the separators of the graph G, and where:

$$P(X_C|\Sigma_C) \propto |\Sigma_C|^{-n/2} exp[-\frac{1}{2} tr\left(S_C\left(\Sigma_C\right)^{-1}\right) \tag{3.6}$$

and similarly for $P(X_S|\Sigma_S)$. Operationally, a model selection procedure compares different G structures by calculating the previous likelihood substituting for Σ its maximum likelihood estimator under G. For a complete (fully connected) Graphical Gaussian Model such an estimator is simply the observed variance-covariance matrix. For a general (decomposable) incomplete graph, an iterative procedure, based on the clique and separators of a graph, must be undertaken (see e.g. [Lauritzen 1996]). Through model selection, we obtain a graphical model that can be used to describe relationships between banks and, specifically, to understand how risks propagate in a systemic risk perspective. [Cerchiello and Giudici 2015] and [Giudici and Spelta 2016] have shown, respectively in the context of country financial flows and bank returns, that Graphical Gaussian models are well suited to estimate interconnections between a large set of financial institutions, on the basis, respectively, of the available inter-country bank liability data or financial market data. In our context, we have the additional task of selecting a graphical model for two different data sources: market data and financial tweets on the same banks. Indeed, the two data sources should be combined into a single one, before performing model selection. This is the additional contribution of the present work, and can be achieved within a Bayesian framework, characterised by an Empirical Bayes approach to the specification of the prior distribution.

Empirical Bayes models [Casella 1985, Carlin and Louis 2000] address the issue of specifying the prior distribution, not on a priori ground, but using data assumed to come from a different population from the one considered as the main object of the statistical inference. In our context, the main object of inference is the correlation structure of market prices, which can be summarised in the correlation matrix parameter. Eliciting a prior distribution on a correlation matrix is a rather complex task, especially when a large number of variables is involved. Furthermore, even when feasible, such a prior may be highly influential on final inferences possibly distorting Bayesian estimates toward the prior, rather than towards the actual data (see e.g. [Casella 1985, Carlin and Louis 2000] and in a financial context, [Giudici

2001]). The Empirical Bayes approach offers a possible solution to this problem. In fact, it allows the prior distribution to be also estimated from real data, possibly different from what used as main object of the inference. In our context, such data is available from Twitter and, therefore, can be employed to estimate an "a priori" correlation matrix. This prior, based on sentiment data, will be then combined with the market price correlation matrix in a Bayesian model. More formally, we first specify a prior distribution for the parameter Σ. [Dawid and Lauritzen 1993] propose a convenient prior, the hyper inverse Wishart distribution. The hyper inverse Wishart distribution can be obtained from a collection of clique-specific marginal inverse Wishart as follows:

$$l(\Sigma) = \frac{\prod_{C \in \mathscr{C}} l(\Sigma_C)}{\prod_{S \in \mathscr{S}} l(\Sigma_S)}, \tag{3.7}$$

where $l(\Sigma_C)$ is the density of an inverse Wishart distribution:

$$l(\Sigma_C) = \frac{|T_C|^{\frac{\alpha}{2}}}{2^{\frac{\alpha p}{2}} \Gamma_p(\frac{\alpha}{2})} |\Sigma_C|^{-\frac{\alpha+p-1}{2}} \exp(-1/2) tr(T_C \Sigma_C^{-1}) \tag{3.8}$$

with hyperparameters T_C and α, and similarly for $l(\Sigma_S)$. For the definition of the hyperparameters here we follow [Giudici and Green 1999] and let T_C and T_S be the submatrices of a larger "scale" matrix T_0 of dimension $N \times N$, and choose $\alpha > N$. [Lauritzen 1996] and [Giudici and Green 1999] show that, under the previous assumptions, the posterior distribution of the variance-covariance matrix Σ is a hyper Wishart distribution with $\alpha + n$ degrees of freedom and a scale matrix given by:

$$T_n = T_0 + S_n \tag{3.9}$$

where S_n is the sample variance-covariance matrix. The previous result can be used to combine market data with tweet data in a Bayesian prior to posterior analysis assuming that the former represent "data" and the latter "prior information" . To achieve this task we recall that, under a complete, fully connected graph, the expected value of the previous inverse Wishart is:

$$E(\Sigma|X) = T_n = (T_0 + S_n)/(\alpha + n) \tag{3.10}$$

and, therefore, the Bayesian estimator of the unknown variance covariance matrix, the a posteriori mean, is a linear combination between the prior (Twitter data) mean and the observed (market data) mean. When the graph G is not complete, a similar result holds locally, at the level of each clique and separator. The previous results suggest to use the above posterior mean as the variance-covariance matrix of a complete graph on which to base model selection. This leads to a new selected graphical model based on a "mixed" data source, containing both financial and tweet data in proportions determined by the quantities α and n. The model selection can be performed by maximizing, rather than the likelihood, the Bayesian a posteriori probability. To achieve this task in an efficient way we will implement a Markov Chain Monte Carlo algorithm, following [Giudici and Green 1999].

We now consider the issue of quantitative learning. In the context of systemic risk, a relevant quantity to be estimated is the partial correlation coefficient which, interpretationally, corresponds to the geometric mean between two regression coefficients in two differently directed multiple regression model. More formally:

$$\rho_{ijV} = \rho_{jiV} = \sqrt{a_{ijV} \cdot a_{jiV}}. \tag{3.11}$$

where a_{ijV} and a_{jiV} are, respectively, the regression coefficient of the multiple regression of X_i on all other V variables (including X_j) and the regression coefficient of the multiple regression of X_j on all other V variables (including X_i).

This interpretation of the partial correlation coefficient helps the construction of a novel contagion effect model. This is based on "modifying" a financial institution's probability of default with the contagion effect from the institutions to which it is connected. The magnitude of this modification is specified by the partial correlation coefficient. For each node (institution) we assume to know the "idiosyncratic" probability of default, π_i. This could be estimated for example on the basis of the rating assigned by a rating agency, or of a credit scoring calculation based on balance sheet data. From the probability of default we can derive, through the inverse Gaussian cumulative distribution function, the (idiosyncratic) credit score of the corresponding institution, as follows:

$$Z_i^0 = \phi^{-1}(1 - \pi_i)$$

where π_i is the default probability of institution i and $1 - \pi_i$ is the corresponding survival probability. We then assume that the idiosyncratic score of an institution can be modified through contagion, in a manner that depends on the credit scores of the neighbours, and on their partial correlations with i, as follows:

$$Z_i^{'} = \phi^{-1}(1 - \pi_0) - \sum_{j \in neigh(i)} a_{ij|rest} \phi^{-1}(1 - \pi_i) \tag{3.12}$$

where $a_{ij|rest}$ is the partial correlation coefficient between variables X_i and X_j given all the others (rest). To interpret the previous assumption, consider the frequent case of positive partial correlations (which occur when banks are highly interrelated, as it occurs within the same country) and negative scores (which occur when default probabilities are less than 50%). In this case the idiosyncratic score increases through contagion and, therefore, the default probability increases too. The modification to the credit score is schematized in Figure 3.1.

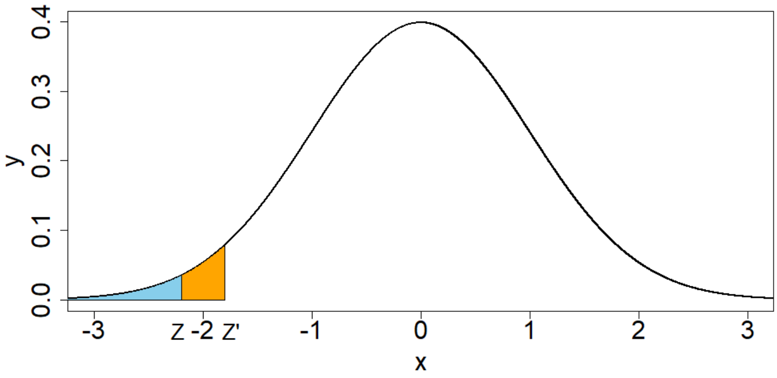


FIGURE 3.1: The impact of contagion on the probability of default: z is the credit score before contagion with the corresponding probability of default coloured in light blue and z' is the credit score after contagion with the corresponding probability of default given by the sum of the light blue and orange area.

3.4 Data

For reasons of information homogeneity we concentrate on a single market: the Italian banking system. The Italian banking system is characterized by a large number of banks operating in a rapidly changing environment due to reforms and the mutated economic conjuncture. We focus on large listed banks, for which daily financial market data exist and can be compared and integrated with tweets data.

Table 3.1 contains the list of the considered banks along with a measure of bank size, their total assets at the end of the last quarter of 2013 (in Euro). Banks are described by their stock market code (ticker).

Bank Name	**Ticker**	**Total Assets**
UniCredit	UCG	926,827
Intesa Sanpaolo	ISP	673,472
Banca Monte dei Paschi di Siena	BMPS	218,882
Unione di Banche Italiane	UBI	132,433
Banco Popolare	BP	131,921
Mediobanca	MB	72,841
Banca Popolare dell'Emilia Romagna	BPER	61,637
Banca Popolare di Milano	BPM	52,475
Banca Carige	CRG	49,325
Banca Popolare di Sondrio	BPSO	32,349
Credito Emiliano	CE	30,748
Credito Valtellinese	CVAL	29,896

TABLE 3.1: List of considered listed Italian Banks

For each bank we calculate the daily stock returns, obtained from the closing price of financial markets, for a period of 148 consecutive days, from July 2013 to February 2014, as follows:

$$R_t = log(P_t/P_{t-1}) \tag{3.13}$$

where t is the day index, $t-1$ the preceding day index and P_t and (P_{t-1}) are the closing prices of the considered bank stock corresponding to the day indexes t and $t-1$.

For the same period, we have crawled Twitter [1] and selected a relevant set of tweets based on their text content to match with the market data. The selected tweets are those containing either one of the banks in Table 3.1 either a keyword belonging to a financial taxonomy. The taxonomy has been developed on the basis of which balance sheet information may affect financial risk; in Table 3.2 we report the taxonomy thematics used to generate the search keywords.

Assets	**Liabilities**	**P& L**
Liquidity	Deposits	Commissions
Corporate bonds	Customer deposits	Interest Margin
Government bonds	Allsale funding	Labour Costs
Loans	Interbank funding	Loans
Consumer loans	Capital	Loans losses
Derivatives	Equity	
	Shares	

TABLE 3.2: Initial proposed taxonomy analysis

[1] We have crawled Twitter using the open source TwitteR package available within the R project environment

Keywords in Table 3.2 have been tested preliminarly to select the most effective in obtaining informative tweets. Keywords contained in Table 3.3 have been regarded as the most relevant figures contained in a bank balance sheet. The complete original taxonomy is longer and more detailed considering synonyms and acronyms as well, but only the few reported in Table 3.3 are characterized by relevant frequencies.

Before extracting tweets, we have preliminarly filtered the most relevant financial twitterers, using the T-index methodology proposed in [Cerchiello and Giudici 2015]. Such methodology relies on an index that ranks sources according to the number of posted tweets, and the corresponding re-tweets obtained. The higher the $T-index$, the stronger is the informative impact of a twitterer because not only she/he posts many tweets but they are also highly shared among the community.

For a formal definition, given a set of n tweets posted by a twitterer to which a retweets count vector of each tweet is associated, we consider the ordered sample of retweets $\{X_{(i)}\}$, that is $X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n)}$, from which $X_{(1)}$ ($X_{(n)}$) denotes the most (the least) cited tweet. Consequently, the T index can be defined as follows:

$$T = max\{t : X_{(t)} \geq t\} \tag{3.14}$$

Once completed the preliminary phase as described above, each selected tweet has been classified into a sentiment class, with scores ranging from 1 to 5. The higher the category, the more positive the sentiment that the tweet assigns to the bank under analysis(1:very negative, 2:negative, 3:neutral, 4:positive and 5:very positive). The sentiment classification has been carried out according to an appropriate classifier, trained on the data and relying on a vocabulary of positive and negative Italian words adapted to the specific financial application under analysis in [Cerchiello and Giudici 2016b]. Such vocabulary is inspired by the famous opinion lexicon (first presented in [Hu M., Liu B. 2004]) that comprises around 6,400 terms. In addition, several experiments and manual cross check have been carried out to improve the reliability and stability of the results. Moreover, since the total number of analyzed tweets is around 1,000, thus easily manageable, the quality of the sentiment classification has been tested accurately comparing methods based on different versions of the vocabulary.

Table 3.3 describes the final employed taxonomy, along with the average sentiment associated to each keyword. Here the sentiment scores are grouped by keywords, so that the average sentiment takes into account all the sentiment scores obtained for that specific word, regardless of the analysed bank.

Keyword	**Frequency*100**	**Average Sentiment**
Commissions	0.03	2.67
Labour costs	1.49	3.21
Deposits	0.08	2.83
Interbank	0.14	2.19
Management	28.58	3.01
Interest margin	4.91	2.79
Subsidiaries	0.99	3.02
Capital	35.67	3.07
Loan losses	0.73	2.90
Loans	10.11	2.93

TABLE 3.3: Taxonomy proposed and descriptive sentiment analysis

For each bank we have then calculated the daily sentiment variation, mimicking the market returns:

$$S_t = log(T_t/T_{t-1}) \tag{3.15}$$

where t is the day index, $t-1$ the preceding day index, and T_t and (T_{t-1}) are the average daily sentiment scores for the considered bank tweets corresponding to the day indexes t and $t-1$.

3.5 Results

In this section we consider the application of our proposed Bayesian model. In terms of prior parameters, we assume that $\alpha = n+2$ and that T is a diagonal matrix, which implies zero a priori partial correlations. Later on we also test the stability of the model to variations of these parameters.

Initially the MCMC procedure described in Section 3.3 is applied to estimate a graphical model from the twelve daily sentiment timeseries obtained from the preprocessing of the tweet data. In this case the prior on the partial correlations is given by the diagonal matrix T. The MCMC based on the Metropolis-Hastings algorithm is run for 500,000 iterations with 10,000 iterations of burn-in. Finally the resulting partial correlation matrix representative of the graphical model obtained from tweet data is used as prior in place of T when estimating the graphical model from the twelve daily stock return timeseries (market data). Also in this case the MCMC algorithm has been run for 500,000 iterations with a burn-in of 10,000 iterations. The resulting partial correlations matrix is representative of both the graphical models estimated on the sentiment timeseries and on the stock returns timeseries with proportion defined by the parameter α. In terms of structural learning, the selected model is the fully connected model: this is quite reasonable, in a national market that is fully integrated, with a strong country effect on bank risk.

Concerning quantitative learning, we report in Table 3.4, below the estimated partial correlations, obtained by model averaging them over the most likely models from the last 10,000 iterations of the MCMC (including, the fully connected model). In Table 3.4 we also report, as a systemic risk measure for each bank, their weighted degree, calculated as the sum of all partial correlations, that expresses the intensity of the contagion.

Table 3.4 and the weighted degree in the last row indicate, which could be the most correlated banks: BPE, BP, followed by BPM and UBI: these are the four largest cooperative banks that are indeed linked to each other. The three largest (public) banks, UCG, ISP and MB, follow. Other smaller banks as well as the troubled MPS have a lesser degree. It is very interesting to notice the high correlation among UCG and ISP which are by far the two largest Italian, which is the highest among all the bank pairs. This is reasonable since both of these banks in similar measure are very influenced by the Italian economy dynamics and by its public debt spread. Table 3.4 is also very useful to draw "stress test" analysis. For example: if UCG returns drop by 100 basis points, each of the other connected banks drop, on average, by 7 basis points, with a total impact on the system of 81 basis point. A similar drop in a smaller and relatively isolated bank, such as CVAL, causes an average drop of the other banks of only 3 basis points.

However the above conclusions do not take bank size into account. It is very likely that the contagion effects among banks depends also on the relative size of their balance sheets. The impact of a large bank, like UCG, on a smaller bank, such as CE, in more extreme scenarios is likely to be greater than what expressed by the weighted degree in Table 3.4. To take size into account, we propose a modification to the calculation of the contagion effect on the probability of default, in Equation 3.12. We introduce a weight that is equal to the ratio of the total assets of the considered bank over the total market assets:

Bank	**UCG**	**UBI**	**MB**	**ISP**	**CVAL**	**CE**	**BP**	**BPSO**	**BPM**	**BPE**	**BMPS**	**CRG**
UCG	1.00	0.01	0.16	0.41	-0.11	-0.04	0.19	0.05	0.09	0.11	0.01	-0.06
UBI	0.01	1.00	0.20	0.03	-0.11	-0.04	0.26	-0.07	0.08	0.26	0.08	-0.03
MB	0.16	0.20	1.00	0.18	0.11	0.10	-0.05	0.10	-0.06	0.05	0.08	0.03
ISP	0.41	0.03	0.18	1.00	-0.09	0.02	-0.00	-0.01	0.13	0.01	0.01	0.00
CVAL	-0.11	-0.11	0.11	-0.09	1.00	-0.01	0.25	0.00	0.07	0.06	0.12	0.06
CE	-0.04	-0.04	0.10	0.02	-0.01	1.00	-0.02	0.09	0.08	0.14	-0.05	-0.04
BP	0.19	0.26	-0.05	0.00	0.25	-0.02	1.00	0.03	0.16	0.23	-0.01	-0.01
BPSO	0.05	-0.07	0.10	-0.01	0.00	0.09	0.03	1.00	0.15	0.04	0.00	0.05
BPM	0.09	0.08	-0.06	0.13	0.07	0.08	0.16	0.015	1.00	0.10	0.14	0.04
BPER	0.11	0.26	0.05	0.01	0.06	0.14	0.23	0.04	0.10	1.00	0.02	0.06
BMPS	0.01	0.08	0.08	0.01	0.12	-0.05	-0.01	0.00	0.14	0.02	1.00	0.11
CRG	-0.06	-0.03	0.03	0.00	0.06	-0.04	-0.01	0.05	0.04	0.06	0.11	1.00
Num. Links	11	11	11	11	11	11	11	11	11	11	11	11
Sum Par. Corr.	0.81	0.90	0.67	0.70	0.35	0.24	1.02	0.43	0.97	1.09	0.51	0.20

TABLE 3.4: Partial correlations and sum of a banks partial correlations with its neighbours based on the selected mixed graphical Gaussian models

$$Z_i' = \phi^{-1}(1-\pi_0) - \sum_{j \in neigh(i)} a_{ij|rest}\,\phi^{-1}(1-\pi_i) * \frac{A_j}{A_{Tot}} \tag{3.16}$$

where A_j represent the assets of the j^{th} neighbour of the bank i (for which the credit score is being calculated) and A_{Tot} are the sum of all the total assets of the different banks.

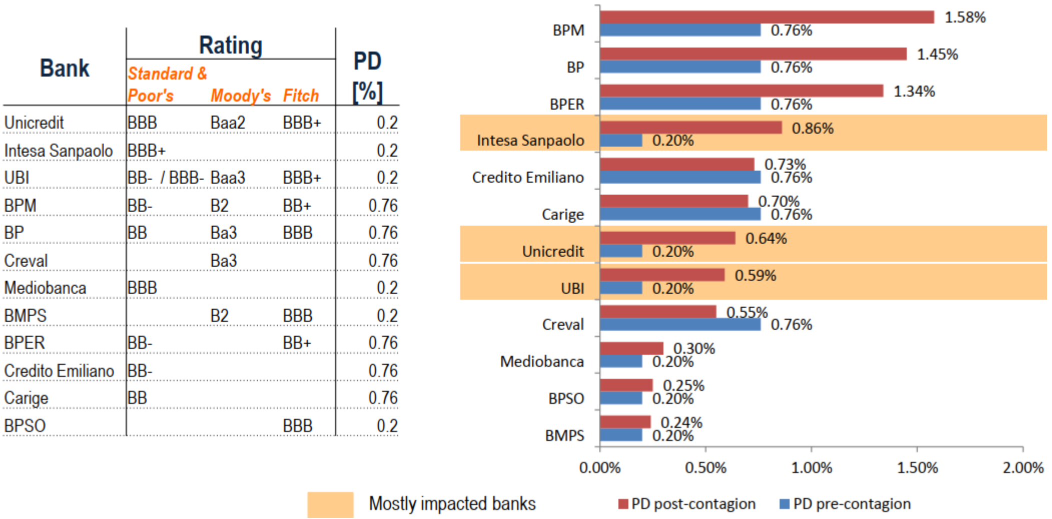
From this equation and an initial estimate of the idiosyncratic probability of default (e.g. that associated with the bank ratings from the rating agencies) we can calculate the banks probability of default incorporating the proposed contagion effect. Accordingly, Table 3.5 and Figure 3.2 indicate the effect of contagion on the idiosyncratic PDs of the considered banks. The second and the third column of the table indicate the probability of default before contagion, and the corresponding percentage variation (ΔPD) with respect to the original probability of default (e.g. +100 corresponds to a 100% relative increase in the probability of default and thus a doubling of the original PD). For robustness purposes, we have also reported the same percentage variation assuming different values for the prior parameter T: a common partial correlation of 0.8, rather than 0, which correspond to a more connected graph in the a priori twitter structure, and different values of the parameter α, which correspond to a higher weight for the twitter prior.

Bank	**Contagioned PD**	$\Delta P.D.$ [%]	$\Delta P.D._{T=0.8}$ [%]	$\Delta P.D._{\alpha=3*(n+2)}$ [%]	$\Delta P.D._{\alpha=30*(n+2)}$ [%]
UCG	0.0064	+220	+220	+220	+220
UBI	0.0059	+195	+175	+200	+200
MB	0.0030	+50	+95	+50	+50
ISP	0.0086	+330	+365	+330	+335
CVAL	0.0055	-28	-1	-28	-28
CE	0.0073	-4	+20	-4	-4
BP	0.0145	+91	+66	+91	+91
BPSO	0.0025	+25	+75	+25	+25
BPM	0.0145	+108	+70	+109	+109
BPER	0.0134	+76	+80	+76	+78
BMPS	0.0024	+20	+5	+20	+20
CRG	0.0070	-8	+12	-8	-8

TABLE 3.5: Partial correlations and systemic risk measures based on the selected mixed graphical Gaussian model

Figure 3.2 illustrate the impact of the contagion effect on the probability of default and how it modifies it. The bank ratings from the rating agencies have been used to estimate the idiosyncratic probability of default. Then applying Equation 3.16 to the partial correlation matrix obtained the graphical model it has been possible to calculate the updated banks PDs taking into account the systemic interconnections. This is a very interesting result because it provide us with a framework to systematically monitor the effect of bank interconnections starting from stock and tweet data.

From both Figure 3.2 and Table 3.5 we can see that the banks which are most impacted by contagion (in relative terms) are the largest banks ISP, UCG as well as UBI, which is the most connected of the cooperative banks. This result finds justification in the fact that due to their size these banks are those that play the most central role in the Italian banking system. Thus, they have on average strong connections with the other banks as can be seen also from the partial correlation matrix in Table 3.5. In terms of robustness analysis, changing the a priori parameters T and α does not change sensibly the results and this indicates stability of the proposed model.



Bank	Rating			PD [%]
	Standard & Poor's	Moody's	Fitch	
Unicredit	BBB	Baa2	BBB+	0.2
Intesa Sanpaolo	BBB+			0.2
UBI	BB- / BBB-	Baa3	BBB+	0.2
BPM	BB-	B2	BB+	0.76
BP	BB	Ba3	BBB	0.76
Creval		Ba3		0.76
Mediobanca	BBB			0.2
BMPS		B2	BBB	0.2
BPER	BB-		BB+	0.76
Credito Emiliano	BB-			0.76
Carige	BB			0.76
BPSO			BBB	0.2

FIGURE 3.2: Impact of the contagion effect on the idiosyncratic probability of default. On the left of the image we can see the bank ratings from the rating agencies that have been used to estimate the idiosyncratic probability of default. The graph on the right shows the idiosyncratic probability of default in blue and the probability of default modified by the contagion effect in blue.

3.6 Conclusions

In this work we have shown how tweet data can be usefully employed in the field of systemic risk modelling by means of Graphical Gaussian Models. We have provided a methodology to combine tweet based systemic risk networks with those obtained from financial market data, using a Bayesian approach and, correspondingly, a Bayesian model selection procedure. This has allowed to develop a framework for systemic risk analysis that integrates these two different, albeit complementary, sources of information.

The proposed model has been applied to the case study of the Italian banking system allowing to estimate the effect of banks linkages starting from tweets on the twelve major Italian banks and their market price data. The developed systemic risk model can be very useful to estimate and take into account contagion risk and its effect on the idiosyncratic probability of default. This kind of analysis can help to individuate vulnerable financial institutions.

Another important value of the model is its capability of including in systemic risk models institutions that are not publicly listed, using the tweet data component alone. This is a relevant advantage for banking systems with many unlisted banks, as it occurs in several European countries, for instance.

The model can be extended in several directions. A promising one could be to replace the inverse cumulative Gaussian link with an extreme value one, as in [Calabrese and Giudici 2015] so to focus more the analysis on tail events.

Chapter 4

Information network modeling for U.S. banking systemic risk

4.1 Summary

In this chapter we investigate whether information theory measures derived from a bank network model, like mutual information and transfer entropy, Granger cause financial stress indexes such as LIBOR-OIS spread, STLFSI and USD/CHF exchange rate. The information theory measures are computed from a Gaussian Graphical Model fitted on the daily stock time series of the top 74 listed US banks. The graphical model is calculated with a recently developed algorithm (LOGO) characterized by a very fast inference that allows us to update the graphical model each market day. From these daily updates of the Graphical Models we can derive daily time series of mutual information and transfer entropy for each bank of the network from April 2003 to May 2017. The Granger causality between the bank related measures and the financial stress indexes is investigated with both standard Granger-causality and Partial Granger-causality conditioned on control measures representative of the general economy stress.

4.2 Introduction

The stability of the financial system is a basic condition for sustainable growth of an economy as a whole. Its importance arises from the key role of financial institutions in capital allocation, i.e. the transfer of financial resources from entities with funds surplus to entities with funds deficit. The 2008 crisis, triggered by large writedowns of bank assets related to subprime mortgages, unfortunately highlighted this principle. This crisis was characterized by the bankruptcy or distress of several large banks like Bear Stearns, Citigroup, Lehman Brothers, Merrill Lynch, Wachovia, and Washington Mutual that in several cases, had to be rescued by the government. This instability of the financial system resulted in a severe credit and liquidity crisis in the financial markets affecting the real economy. This type of risk, wherein the entire financial system is simultaneously distressed, is generally referred to as systemic risk. Systemic risk, when realized, impacts not only financial markets and institutions, but also the real economy as a whole due to decreases in capital supply and increases in capital costs.

The term systemic risk was coined the early 1980s by the economist William Cline [Ozgöde 2011] at the onset of the Latin American debt crisis. According to his definition, systemic risk is a threat that disturbances in the financial system will have serious adverse effects on the entire financial market and the real economy. Systemic risk models address the issue of interdependence between financial institutions and, specifically, measure how bank default risks are transmitted among banks.

The last few years have witnessed an increasing research literature on systemic risk, with the aim of identifying the most contagious institutions and their transmission channels. Specific measures of systemic risk have been proposed for the banking sector; in particular by [Adrian and Brunnermeier 2009], [Acharya et al. 2010](MES), [Brownlees and Engle 2011, Huang et al. 2011], [Acharya et al. 2012](SRISK), [Cao 2013](δCoVaR), [Banulescu and Dumitrescu 2015](CES), [Calabrese and Giudici 2015]. These approaches leverage financial market price information to asses the financial institution's appropriate quantiles of the estimated loss probability distribution, conditional on a crash event in the financial market. However, they do not address the issue of risk transmission between different banks. In order to, address this aspect of systemic risk, researchers have introduced financial network models. Networks have emerged as a useful tool for understanding contagion and systemic risk, in financial systems. In fact, after the 2008 financial crisis, there have been many studies on financial networks and their role in systemic risk. A major finding emphasized by these studies is that financial contagion is mainly driven by system-wide interconnectedness of institutions. In particular, [Billio et al. 2012] propose several econometric measures of connectedness based on Granger-causality networks and principal component analysis. [Hautsch et al. 2014, Peltonen et al. 2015] propose tail dependence network models aimed at overcoming the bivariate nature of the available systemic risk measures. [Diebold and Yilmaz 2014] propose LASSO regularized Vector Autoregressive models for selecting the significant links in a network model. Network models are based on the assumption of full connectedness among all nodes, which makes their interpretation difficult and also estimation when a large number of them is being considered. Trying to tackle this limitation, [Giudici and Spelta 2016] and [Cerchiello and Giudici 2016] resorted to graphical correlation models, which can account for partial connectedness, expressed in terms of conditional independence constraints. A similar but alternative approach has been explored by [Barigozzi et al. 2013] introducing multivariate Brownian processes with a correlation structure determined by a conditional independence graph.

Correlation networks have proven a suitable tool to visualize the structure of pairwise marginal correlations among a set of nodes corresponding to the investigated banking systems. In these models each banks is represented by a node in the network, and each pair of nodes can be connected by an edge, which has a weight related to the correlation coefficient between the two nodes. Furthermore, the banking system represented with these models can be described by the adjacency and inverse covariance matrix of the corresponding graphical model.

Our contribution follows this latter development estimating a Graphical Gaussian Model on the market prices of the 74 largest listed U.S. banks. We estimate the model with a recently developed algorithm (LoGo) for reconstructing the sparse inverse covariance matrix from the data.

The LOGO algorithm is characterized by a very fast inference that allows us to update the graphical model for each market day of the observation period. With these daily updates of the Graphical Models we can generate daily time series of the mutual information and the transfer entropy for the entire system and also for each bank of the network, from April 2003 to May 2017.

Then we investigate how the information theory measures (mutual information and transfer entropy) derived from the graphical model, correlate with and Granger cause financial stress indexes like LIBOR-OIS spread, STLFSI and USD/CHF exchange rate. The idea is to understand how these measures compare to the financial stress indexes and which banks show Granger causality links with the indexes. The Granger causality between the bank related measures and the financial stress indexes is investigated applying Partial Granger-causality tests conditioning on control measures representative of the economic and financial system stress.

The remainder of this chapter is organized as follows. In Section 4.3 we present in different subsections the theory and the models applied in this study. Subsection 4.3.1 introduces the graphical models and their theoretical background. In Subsection 4.3.2 we briefly recall the LoGo methodology that we use to infer the graphical model on the bank network. In Subsection 4.3.3 we introduce the measures that we calculate from the bank network model. In Subsections 4.3.4 and 4.3.5 we presents formally the Granger causality and the partial Granger causality discussing their application. In Section 4.4 we present the bank stocks data that we use to fit the network model and the financial stress indexes whose causality relationship is investigated. In Section 4.5 we present the results of the causality analysis between the measures derived from the bank stock network model and the financial stress indexes. Finally in Section 4.6 we briefly discuss the results of the work.

4.3 Methodology

4.3.1 Network model

Here, we briefly describe the Gaussian graphical models that will be applied to estimate relationships between the N banks, both with market and sentiment data. Direct relationships between banks can be measured by their partial correlation, that expresses the direct influence of a bank on another. Partial correlations can be estimated assuming that the observations follow a graphical Gaussian model, in which the covariance matrix Σ is constrained by the conditional independences described by a graph (see e.g. [Lauritzen 1996]). More formally, let $X = (X_1,...,X_N) \in R^N$ be a $N-$dimensional random vector distributed according to a multivariate normal distribution $\mathcal{N}(\mu,\Sigma)$. Without loss of generality, we will assume that data are generated by a stationary process, and, therefore, $\mu = 0$. In addition, we will assume throughout that the covariance matrix Σ is not singular.

Let $G = (V,E)$ be an undirected graph, with vertex set $V = \{1,...,N\}$, and edge set $E = V \times V$, a binary matrix, with elements e_{ij}, that describe whether pairs of vertices are (symmetrically) linked between each other ($e_{ij} = 1$), or not ($e_{ij} = 0$). If the vertices V of this graph are put in correspondence with the random variables $(X_1,...,X_N)$, the edge set E induces conditional independence on X via the so-called Markov properties [Lauritzen 1996]. In particular, the pairwise Markov property determined by G states that, for all $1 \leq i < j \leq N$:

$$e_{ij} = 0 \Longleftrightarrow X_i \perp X_j | X_{V\setminus\{i,j\}}; \tag{4.1}$$

that is, the absence of an edge between vertices i and j is equivalent to independence between the random variables X_i and X_j, conditionally on all other variables $x_{V\setminus\{i,j\}}$.

Let the elements of Σ^{-1}, the inverse of the covariance matrix, be indicated as $\{\sigma^{ij}\}$. [Whittaker 1990] proved that the following equivalence also holds:

$$X_i \perp X_j | X_{V\setminus\{i,j\}} \Longleftrightarrow \rho_{ijV} = 0 \tag{4.2}$$

where

$$\rho_{ijV} = \frac{-\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}} \tag{4.3}$$

denotes the ij-th partial correlation, that is, the correlation between X_i and X_j, conditionally on the remaining variables $X_{V\setminus\{i,j\}}$.

Therefore, by means of the pairwise Markov property, and given an undirected graph $G = (V,E)$, a graphical Gaussian model can be defined as the family of all N-variate normal distributions that satisfy the constraints induced by the graph on the partial correlations, as follows:

$$e_{ij} = 0 \Longleftrightarrow \rho_{ijV} = 0 \tag{4.4}$$

for all $1 \leq i < j \leq N$.

4.3.2 LoGo algorithm

In our study we investigate a relatively large number of banks (74) and we take advantage of a recently presented algorithm LoGo [Barfuss et al. 2016] to estimate graphical models on the basis of time series data. LoGo is a methodology that makes use of information filtering networks to produce probabilistic models that are sparse and with high likelihood. One of its main advantages is that it's computationally fast, making possible applications with very large data sets. The LoGo algoritm calculates the global sparse inverse covariance matrix from a simple sum of local inverse covariances computed on small subparts of the network matrices. The use of low-dimensional local inversions makes the procedure computationally efficient, statistically robust and only slightly sensitive to the curse of dimensionality [Barfuss et al. 2016]. In particular the method is based on a recent, new family of information filtering networks, the triangulated maximal planar graph (TMFG) [Massara et al. 2016] that are decomposable graphs. A decomposable graph has the property that every cycle of length greater than three has a chord, an edge that connects two vertices of the cycle in a smaller cycle of length three. The construction of the algorithm, through a sum of local inversion, makes this methodology particularly suitable for parallel computing and dynamical adaptation by local, partial updating, as described in [Barfuss et al. 2016] where a more detailed explanation of the method is presented.

4.3.3 Information theory measures

We leverage the network model fitted on the bank stocks to calculate several bank and system related measures. We fit a new Gaussian Graphical Model for every market day, based on the stock time series of the 90 previous market days, so we can calculate the bank and system related measures every market day and they are representative of the last 90 days trends. We obtain a time series that goes from March 2003 to October 2017 for every measure that we calculate from the network model. We selected a time frame of 90 days to fit the graphical model for three main reasons: i. have more datapoints (90) than the number of banks (74); ii. Obtain a network representative only of the last few months; iii. the LoGo algorithm outperforms the Glasso specially in the case when the number of variables (banks) and datapoint (days) are comparable [Barfuss et al. 2016]. While it's possible to extract many interesting indicators from the network model like bank and system average partial correlations, number of edges, Pagerank and others, we focus on two measures derived from information theory: Mutual Information and Transfer Entropy between banks.

The mutual information is a measure derived from information theory and probability theory that can be calculated among two random variables. It is a measure of the mutual dependence between the two variables and it quantifies the amount of information that one can tell us about the other. Intuitively, it measures the information shared by the two variables and quantifies how much knowing one variable reduces the uncertainty about the other [Mackay 2003]. When two variables are independent, knowing one does not give any information about the other and vice versa, so their mutual information is zero. At the other extreme, if one is a deterministic function of the other and vice versa then all the information conveyed by one variable is shared with the other: knowing just one of them determines the value of the other and vice versa. As a result, in this case the mutual information is the same as the uncertainty contained in one of the variables alone, namely their entropy. In general, if we represent the different entropies of the two random variables with an analogy to set theory, the mutual information is the intersection of the two sets and represents the uncertainty that is common to both the random variables.

The definition of the mutual information for two continuous random variables is:

$$I(X;Y) = \int_Y \int_X p(x,y) log(\frac{p(x,y)}{p(x)p(y)}) dxdy \tag{4.5}$$

Mutual information however says little about causal relationships, because it lacks directional and dynamical information. In fact, it is symmetric between the random variables and thus, it cannot distinguish between driver and response variables [Vicente et al. 2011].

Also the transfer entropy is a measure derived from information theory and probability theory that can be calculated among two random variables. It is a measure of the amount of directed (time-asymmetric) transfer of information between the two variables. Thus, transfer entropy from a random variable X to another random variable Y is the amount of uncertainty reduced in future values of Y by knowing the past values of X given past values of Y [Schreiber et al. 2000]. In other words, the Transfer entropy is the conditional mutual information, with the history of the influenced variable $Y_{t-1:t-L}$ in the condition [Wyner 1978]

$$T_{X \to Y} = I(Y_t; X_{t-1:t-L} \mid Y_{t-1:t-L}). \tag{4.6}$$

This means that the transfer entropy can be taken as an indicator to understand which are the driver and response variables in a system [Schindler et al. 2007].

In our study we calculate the mutual information and the transfer entropy among the banks of the network for each day of the observation period to produce a corresponding time series for each bank. The mutual information is updated every market day for every edge of the network (GGM inferred by the LoGo algorithm) and then for each bank we take the summation of the mutual information over its network edges. In other words, we take summation of mutual information between a bank and all its first neighbours. As a result we obtain a timeseries for each bank that describes the evolution of the total mutual information of the bank with its neighbours.

Similarly we obtain a time series for the transfer entropy but in order to calculate the transfer entropy we need the time dimension within the model so we had to fit the GGM with the LoGo algorithm considering also lag-1 variables, since the transfer entropy exists between lagged and contemporary variables as shown in the formula 4.6. So, every bank in this network is represented by a contemporary variable (e.g. JPM_t) and a lag-1 variable (e.g. $JPM_t - 1$). For every bank couple (e.g. JPM and BAC) we have two "transfer entropies": $JPM_t - 1$ to BAC_t and vice versa $BAC_t - 1$ to JPM_t; the first ($JPM_t - 1$ to BAC_t) is a transfer entropy inflow for BAC and a transfer entropy outflow for JPM while the second ($BAC_t - 1$ to JPM_t) is a transfer entropy inflow for JPM and transfer entropy outflow for BAC. Thus, we have two transfer entropy measures for each bank: "transfer entropy inflow" and "transfer entropy outflow". As for the mutual information, for each bank we take the summation of these quantities over its edges (the bank first neighbours). The transfer entropy inflow is related to how much a bank stock behaviour is predictable given the previous behaviour of its neighbours while the outflow is related to how much a bank stock previous behaviour is useful for predicting its neighbours stocks.

4.3.4 Granger Causality

Since our research hypothesis aims at analyzing whether information theory measures extracted from the bank network are useful to predict the financial stress indexes and if they cause them according to a temporal dimension, we need a method to assess such effect. In this paragraph, we introduce the Granger causality test, a well-known econometric test useful when causality is the object of interest. Granger causality entails the statistical notion of causality based on the relative forecast power of two time series. Time series j is said to “Granger-cause” time series i if past values of j contain information that helps in predicting i above and beyond the information contained in past values of i alone. In a well known paper [Granger 1969], Granger has proposed a useful test based on the following principle: if lagged values of time series X_t contribute to foresee current values of time series Y_t in a forecast achieved with lagged values of both X_t and Y_t, then we say X_t *Granger causes* Y_t. As was first shown in [Sims 1972], the Granger causality corresponds to the concept of exogeneity and it is therefore necessary to have a unidirectional causality in order to guarantee consistent estimation of distributed lag models. The mathematical formulation of this test is based on linear regressions of X_{t+1} on X_t and Y_t

In our research framework, we propose to calculate the Granger causality test on pairs of times series defined as follows:

- R_t^{kq}: given a network measure k, for bank q at time t.
- R_t^p: given a financial stress index p at time t.

To ease the notation we refer to R_t^q given the network measure k and bank q. Thus, applying the test for a given pair of network measures for bank q and financial stress index p, we result in fitting the following equations:

$$R_{t+i}^p = \beta_0^p R_t^p + \beta_1^{pq} R_t^q + e_{t+i}^p \tag{4.7}$$

$$R_{t+i}^q = \beta_0^q R_t^q + \beta_1^{qp} R_t^p + e_{t+i}^q \tag{4.8}$$

Our null hypothesis is therefore: $H_0 : \beta_1^{pq} = \beta_1^{qp} = 0$. Taking into account that we are dealing with daily time series, in our tests we have considered up to five lags as plausible windows of analysis to be able to represent the effects of one business week.

4.3.5 Linear Partial Granger Causality

Often economic network models, like in our case, involve the step of 'structural model selection', in which a relevant set of variables is selected for analysis. In practice, this step is likely to exclude some relevant variables, which can lead to the detection of apparent causal interactions that are actually spurious [Pearl 1999]. One recent response to this challenge has been the 'partial Granger causality' method introduced in [Guo et al. 2008]. The idea is that latent variables may give rise to detectable correlations among the residuals of the corresponding vector autoregressive model. By analogy with the concept of partial correlation [Kendall and Stuart 1979], an additional term based on these correlations can mitigate the confounding influence of the latent variables. Later, [Barrett et al. 2010], redefined the F1 statistic as a model comparison between nested models where the present conditioning variables are added as predictors to the autoregressive model.

The linear partial Granger causality is defined as follows. Consider two time series X_t and Z_t which admit a joint autoregressive representation of the form

$$X_t = \sum_{i=1}^{\infty} a_{1i} X_{t-i} + \sum_{i=1}^{\infty} c_{1i} Z_{t-i} + \overrightarrow{\varepsilon_{1t}} + \overrightarrow{\varepsilon_{1t}^E} + \overrightarrow{B_1(L)\varepsilon_{1t}^L} \tag{4.9}$$

$$Z_t = \sum_{i=1}^{\infty} b_{1i} Z_{t-i} + \sum_{i=1}^{\infty} d_{1i} X_{t-i} + \overrightarrow{\varepsilon_{2t}} + \overrightarrow{\varepsilon_{2t}^E} + \overrightarrow{B_2(L)\varepsilon_{2t}^L} \tag{4.10}$$

For simplicity of notation, let us define

$$u_i(t) = \overrightarrow{\varepsilon_{it}} + \overrightarrow{\varepsilon_{it}^{E}} + \overrightarrow{B_i(L)\varepsilon_{it}^{L}} \tag{4.11}$$

where $i = 1,2$. The noise covariance matrix for the model then can be represented as

$$S = \begin{bmatrix} var(u_{1t}) & cov(u_{1t},u_{2t}) \\ cov(u_{2t},u_{1t}) & var(u_{2t}) \end{bmatrix} \tag{4.12}$$

Extending this concept, the vector autoregressive representation for a system involving three variables X_t, Y_t and Z_t can be written as follows:

$$X_t = \sum_{i=1}^{\infty} a_{2i}X_{t-i} + \sum_{i=1}^{\infty} b_{2i}Y_{t-i} + \sum_{i=1}^{\infty} c_{2i}Z_{t-i} + \overrightarrow{\varepsilon_{3t}} + \overrightarrow{\varepsilon_{3t}^{E}} + \overrightarrow{B_3(L)\varepsilon_{3t}^{L}} \tag{4.13}$$

$$Y_t = \sum_{i=1}^{\infty} d_{2i}X_{t-i} + \sum_{i=1}^{\infty} e_{2i}Y_{t-i} + \sum_{i=1}^{\infty} f_{2i}Z_{t-i} + \overrightarrow{\varepsilon_{4t}} + \overrightarrow{\varepsilon_{4t}^{E}} + \overrightarrow{B_4(L)\varepsilon_{4t}^{L}} \tag{4.14}$$

$$Z_t = \sum_{i=1}^{\infty} g_{2i}X_{t-i} + \sum_{i=1}^{\infty} h_{2i}Y_{t-i} + \sum_{i=1}^{\infty} k_{2i}Z_{t-i} + \overrightarrow{\varepsilon_{5t}} + \overrightarrow{\varepsilon_{5t}^{E}} + \overrightarrow{B_5(L)\varepsilon_{5t}^{L}} \tag{4.15}$$

The noise covariance matrix for the model can be represented as

$$\Sigma = \begin{bmatrix} var(u_{3t}) & cov(u_{3t},u_{4t}) & cov(u_{3t},u_{5t}) \\ cov(u_{4t},u_{3t}) & var(u_{4t}) & cov(u_{4t},u_{5t}) \\ cov(u_{5t},u_{3t}) & cov(u_{5t},u_{4t}) & var(u_{5t}) \end{bmatrix} \tag{4.16}$$

where

$$u_i(t) = \overrightarrow{\varepsilon_{it}} + \overrightarrow{\varepsilon_{it}^{E}} + \overrightarrow{B_i(L)\varepsilon_{it}^{L}} \tag{4.17}$$

where $i = 3,4,5$

In order to consider the influence from Y to X controlling for the effect of the exogenous input, the noise covariance matrix S is partitioned in the following way:

$$S = \begin{bmatrix} \frac{var(u_{1t})|cov(u_{1t},u_{2t})}{cov(u_{2t},u_{1t})|var(u_{2t})} \end{bmatrix} = \begin{bmatrix} \frac{S_{11}|S_{12}}{S_{21}|S_{22}} \end{bmatrix} \tag{4.18}$$

Hence the variance of u_{1t} by eliminating the influence of u_{2t} is given by:

$$R_{XX|Z}^{1} = cov(u_{1t},u_{1t}) - cov(u_{1t},u_{2t})cov(u_{2t},u_{2t})^{-1}cov(u_{2t},u_{1t}) = S_{11} - S_{12}S_{22}^{-1}S_{21} \tag{4.19}$$

For the matrix Σ, by eliminating the second row and the second column, the remaining noise covariance matrix Σ can be partitioned in the following way

$$\Sigma = \begin{bmatrix} \frac{var(u_{3t})|cov(u_{3t},u_{5t})}{cov(u_{5t},u_{3t})|var(u_{5t})} \end{bmatrix} = \begin{bmatrix} \frac{\Sigma_{11}|\Sigma_{12}}{\Sigma_{21}|\Sigma_{22}} \end{bmatrix} \tag{4.20}$$

The variance of u_{3t} can be defined by eliminating the influence of u_{5t} similarly

$$R_{XX|Z}^{(2)} = cov(u_{3t},u_{3t}) - cov(u_{3t},u_{5t})cov(u_{5t},u_{5t})^{-1}cov(u_{5t},u_{3t}) = \Sigma_{11} - \Sigma_{12}S_{22}^{-1}\Sigma_{21} \tag{4.21}$$

The value of $R_{XX|Z}^{(1)}$ measures the accuracy of the autoregressive prediction of X based on its previous values conditioned on Z by eliminating the influence of the common exogenous input and latent variables, whereas the value of $R_{XX|Z}^{(2)}$ represents the accuracy of predicting the present value of X based on the previous history of both X and Y conditioned on Z by eliminating the influence of the exogenous input and latent variables. According to the causality definition of Granger, if the prediction of one process is improved by incorporating the information of the second process, then the second process Granger causes the first process. Similarly it is possible to define this causal influence by

$$F_1 = ln\left(\frac{|R^{(1)}_{XX|Z}|}{|R^{(2)}_{XX|Z}|}\right) = ln\left(\frac{S_{11} - S_{12}S_{22}^{-1}S_{21}}{\Sigma_{11} - \Sigma_{12}S_{22}^{-1}\Sigma_{21}}\right) \tag{4.22}$$

F_1 it's called partial Granger causality. For comparison, the standard conditional Granger causality F_2 is defined by

$$F_2 = ln\left(\frac{|S_{11}|}{|\Sigma_{11}|}\right) \tag{4.23}$$

While in theory, partial Granger causality is only able to eliminate confounders effects when their influence is identical for every time series however in [Roelstraete et al. 2012] it has been shown to be robust for deviations from this assumption. Moreover, in the presence of unknown latent and exogenous influences, it is shown in [Guo et al. 2008] and again in [Roelstraete et al. 2012] that partial Granger causality better eliminates their influence than conditional Granger causality and simple Granger causality outperforming both of them.

4.4 Data

The data we analyze are banks stock price time series and for sake of comparability and homogeneity, we focus on a single banking market, the U.S. banking system. This is an interesting group of banks to study, due to its relevance in the world economy and particularly for its role in originating the 2008 financial crisis, with many large banks which have seriously impacted the world and U.S. economy and politics. We take into account the top 74 U.S. large listed banks, for which there exist daily financial market data that we collect. In Table 4.1 we reported the list of the banks that we consider, along with their stock market code (ticker) and their total assets at the end of 2016 (in US dollars).

For each bank, we consider the daily log-returns obtained from the stock closing price of financial markets, for a period of 3,716 days from Jan 2003 through October 2017, as follows:

$$R_t = ln(P_t/P_{t-1}) \tag{4.24}$$

where t is a day, $t-1$ the day preceding it and P_t the corresponding closing price of that bank in that day while P_{t-1} is the closing price of the previous market day.

In our study we want to inspect the causality relations among bank stocks and the overall system financial stress, thus we need to consider some suitable stress indicators. For this we select three three indexes commonly considered when evaluating the stress of the financial system: the St. Louis Fed Financial Stress Index (STLFSI), the London Interbank Offering Rate–Overnight Index Swap spread (LIBOR-OIS spread) and USD/CHF exchange ratio.

The STLFSI is a financial stress index constructed by the Federal Reserve Bank of St. Louis. It measures the degree of financial stress in the markets and is constructed from 18 weekly data series: seven interest rate series, six yield spreads and five other indicators. Each of these variables captures some aspect of the financial stress. Accordingly, as the level of financial stress in the economy changes, the data series are likely to move together. The average value of the index, which begins in late 1993, is designed to be zero. Thus, zero is viewed as representing normal financial market conditions. Values below zero suggest

Bank	Ticker	Assets ($ bn)	Bank	Ticker	Assets ($ bn)
JPMorgan Chase Bank	JPM	2,118	Banco Popular de Puerto Rico	BPOP	30
Wells Fargo Bank	WFC	1,741	Frost Bank	CFR	30
Bank of America	BAC	1,660	Synovus Bank	SNV	29
Citibank	C	1,356	Associated Bank	ASB	29
U.S. Bank	USB	448	First Tennessee Bank	FHN	28
PNC Bank	PNC	358	Webster Bank	WBS	26
The Bank of New York Mellon	BK	300	Umpqua Bank	UMPQ	25
Capital One	COF	279	Commerce Bank	CBSH	25
TD Bank	TD	265	Whitney Bank	HBHC	23
State Street Bank	STT	252	Valley National Bank	VLY	22
Branch Banking and Trust Company	BBT	217	First National Bank of Pennsylvania	FNB	21
HSBC Bank USA	HSBC	204	Prosperity Bank	PB	21
SunTrust Bank	STI	200	Pacific Western Bank	PACW	21
Charles Schwab Bank	SCHW	165	TCF National Bank	TCF	21
Goldman Sachs Bank USA	GS	158	Iberiabank	IBKC	21
Fifth Third Bank	FITB	141	UMB Bank	UMBF	19
Morgan Stanley Bank	MS	127	MB Financial Bank	MBFI	19
Manufacturers and Traders Trust	MTB	126	Bank of the Ozarks	OZRK	18
Regions Bank	RF	124	Sallie Mae Bank	SLM	18
The Northern Trust Company	NTRS	120	Raymond James Bank	RJF	17
MUFG Union Bank	MTU	117	FirstBank	FBP	17
BMO Harris Bank	BMO	107	Bank of Hawaii	BOH	16
KeyBank	KEY	101	Washington Federal	WAFD	15
Huntington National Bank	HBAN	100	Astoria Bank	AF	15
Santander Bank	SAN	85	Old National Bank	ONB	15
Compass Bank	BBVA	85	BancorpSouth Bank	BXS	15
Comerica Bank	CMA	74	Flagstar Bank, FSB	FBC	14
Deutsche Bank Trust Company Americas	DB	55	Cathay Bank	CATY	14
American Express Bank	AXP	46	Sterling National Bank	STL	14
New York Community Bank	NYCB	46	Bank of Hope	HOPE	14
Silicon Valley Bank	SIVB	43	Trustmark National Bank	TRMK	13
People’s United Bank	PBCT	40	First Midwest Bank	FMBI	11
E*TRADE Bank	ETFC	36	Stifel Bank and Trust	SF	11
East West Bank	EWBC	33	Banc of California	BANC	11
First-Citizens Bank & Trust Company	FCNCA	33	Fulton Bank	FULT	11
BOK Financial	BOKF	33	United Community Bank	UCBI	10
Barclays Bank Delaware	BCS	31	State Bank of India	SBIN	10

TABLE 4.1: List of the banks object of the study

below-average financial market stress, while values above zero suggest above-average financial market stress [Federal Reserve 2010, Federal Reserve Bank of St. Louis 2018].

The LIBOR-OIS spread is the difference between the 3-month London Interbank Offered Rate (LIBOR) and the corresponding overnight indexed swap (OIS) rates and is regarded as a strong indicator of the health of the banking system [Sengupta et al. 2008]. The LIBOR is the interest rate at which banks borrow unsecured funds from other banks in the London wholesale money market for a period of 3 months. Alternatively, a bank can enter into an overnight indexed swap (OIS) that entitles it to receive a fixed rate of interest on a notional amount called the OIS rate. In exchange, the bank agrees to pay a (compound) interest payment on the OIS rate to be determined by a reference floating rate (in the United States, this is the effective federal funds rate) to the counterparty at maturity [Sengupta et al. 2008]. Briefly, the LIBOR-OIS spread it's a measure of how expensive or cheap it will be for banks to borrow, as shown by LIBOR, relative to a risk-free rate. It is an important measure of risk and liquidity in the money market, and thus an indicator for the relative stress in the money markets. In general, when the spread is wider (higher LIBOR) is considered as a lower availability to lend by major banks, while a narrow spread indicates higher liquidity in the market. For this reason, the spread can be regarded as an indication of banks' perception of the creditworthiness of the other financial institutions and of the general availability of funds for lending purposes. Compared to LIBOR the LIBOR-OIS spread provides a more complete picture of how the market is viewing credit conditions because it strips out the effects of underlying interest-rate moves, which are in turn affected by factors such as central bank policy, inflation and growth expectations. During the financial crisis of 2007-2010, the LIBOR-OIS spread reached its maximum indicating a severe credit crunch and peaked concurrently with announcements of emergency funding to rescue Northern Rock, large write-downs by large investment banks and large bank failures.

The USD/CHF exchange rate is considered a measure of financial stress because in period of financial stress and instability safe haven inflows are likely to play a key role in the appreciation of the Swiss franc [Deutsche Bundesbank 2014]. Currencies in fact, can appreciate in times of crisis because they are offered as safe investment instruments by the countries issuing them. The currencies of such countries are commonly referred to as safe haven currencies and the media and the literature are unanimous in ascribing the strength of the Swiss franc to its status as a safe haven currency. Empirical findings support this theory like when in mid-2011, the franc appreciated so strongly against the Euro that it almost attained parity, the Swiss National Bank announced that it would defend a minimum exchange rate of 1.20 CHF against the Euro and it was prepared to purchase unlimited amounts of foreign currency if needed [Swiss National Bank 2011]. The aim, as explained in its press release, was to counteract the massive overvaluation of the Swiss franc and protect Swiss economy that is heavily reliant on the exports of goods and services worth over ca. 65% of the GDP [Worldbank 2018].

In our study thus, we investigate the causality relationship among these three financial stress indicators and the measures extracted from the network model inferred with the LoGo algorithm from the bank stock time series. Moreover, when testing the Partial Granger Causality we conditioned on three control variables related to the general economic conjuncture: US 10Y Treasuries yield, gold price and EUR/USD exchange ratio. We derived the time series of the daily returns for both the gold price and the EUR/USD exchange ratio (like for the case of the bank stocks) given by Equation 4.24 while the 10Y US Treasuries yield time series hasn't been preprocessed since it represents an interest rate. We decided to condition on these variables in order to control for the effect of the general status of the economy in the

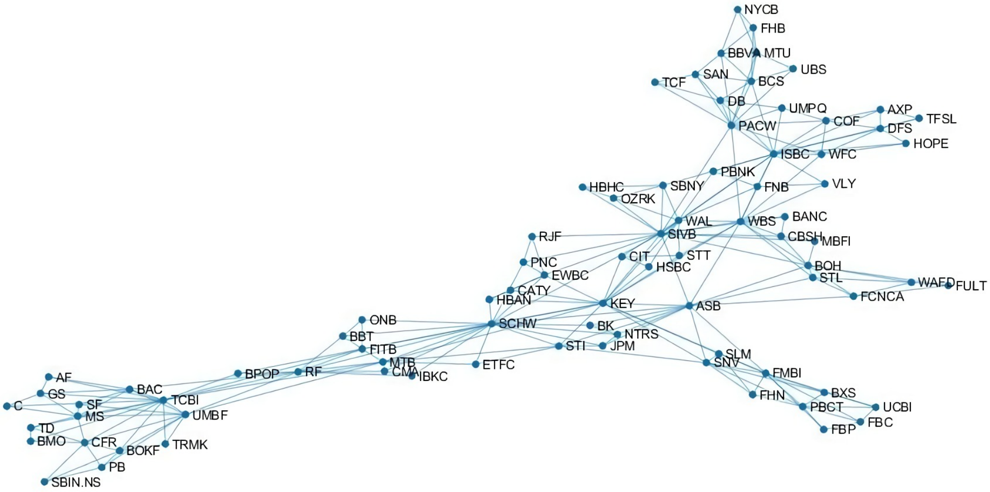


FIGURE 4.1: Network model inferred from stock returns data for the period 2003-2017 by the LoGo algorithm

partial Granger causality analysis.

4.5 Results

Initially, we estimated a graphical model of the U.S. banks over the entire time horizon (2003-2017) with the LoGo algorithm to gather insights regarding the most correlated banks (from the stock price point of view). In this case we have a single time series for each bank with the daily returns of stock closing price spanning from January 2003 to May 2017. Thus, the graphical model obtained is representative of the partial correlations of the returns from 2003 to 2017. Given such a long time frame we would expect to see only some constant properties of the banks emerge from the graphical model structure, like characteristics connected to the bank dimension, business model, nationality. Interestingly from Figure 4.1 we can see that the estimated network posits many of the largest bank like C, BAC, GS, MS, TD close to each other and connected by edges in the lower left corner of the network. At the same time, foreign banks like BBVA, BCS, DB, UBS are located together in the right top corner.

Secondly, we calculate a different graphical model for each market day based on the data of the 90 previous days. Literally, we apply a moving window of length 90 to the stock returns time series and for every step of the moving window we fit a graphical model through the LoGo algorithm. Thus, for every market day we obtain a network representative of the bank stocks returns correlations and market structure in the previous 90 days. All these networks can be imagined as a daily time series of graphical models from May 2003 to May 2017 (we start from May 2003 instead of January, due to the moving window lag). We will leverage these dynamic 'snapshots' of how the U.S. bank system stock correlations evolve to generate several time series of measures derived from the network models. It is possible to compute different bank related measures from the graphical models (namely mutual information, pagerank, transfer entropy from lagged variables, number of bank edges) that allow for an interesting inspection of the system evolution, highlighting different aspects. In order to be able to calculate the transfer entropy, we fit also a model that comprehends 1-day lagged returns as input variables. Thus, the input are two time series of 90 days of length for each bank, one with contemporary returns and one with 1-day lagged returns. When we derive the mutual information time series we don't need to use the model with 1-day lagged

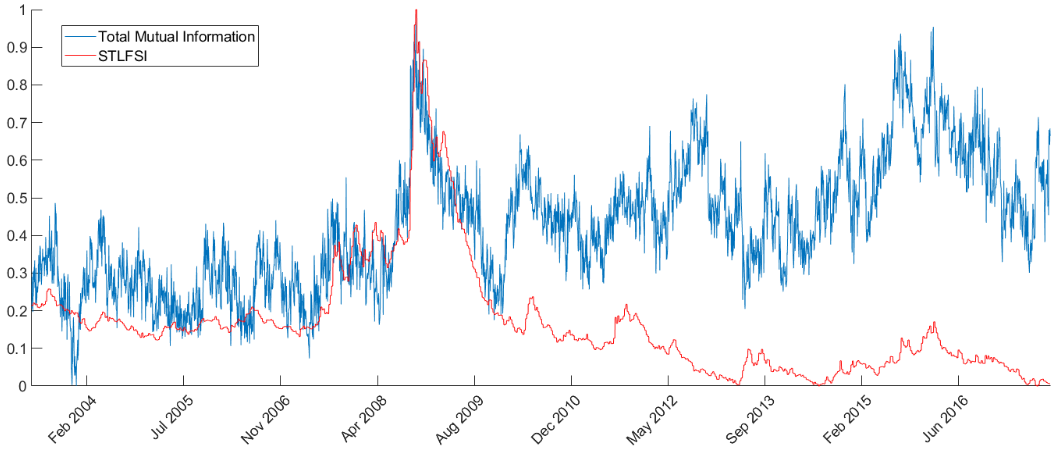


FIGURE 4.2: Total network mutual information vs STLFSI trends comparison

returns, while when computing the transfer entropy time series we use it. Both the mutual information and the transfer entropy are calculated for each edge of every graphical model with the distinction that for transfer entropy is calculated only for the edges that go from 1-day lagged nodes to contemporary ones. Calculating these measures only for the network edges is a great computational saving because it means computing around 100 measures per graphical model (the number of edges of our sparse LoGo inferred model on average) instead of calculating 2,485 quantities ($(n^2 - n)/2$, with n number of nodes). Then we aggregate the mutual information and transfer entropy time series both at bank level and system level. The system level aggregation produces measures that summarize the behaviour of the entire bank network and can be compared with the overall financial system stress indexes. For example, from Figure 4.2 it is possible to see how the network total mutual information resembles very closely (especially in the trends) the STLFSI. We can see from the figure that the trends are very similar and timely coincident, specially around the stress peaks registered during the 2008 financial crisis. This result is coherent with [Dilip et al. 2013] where the authors show that correlation spikes tend to predict or coincide with significant economic or market events, especially during the 2007-2008 financial crisis.

The Transfer entropy trend compared to the STLFSI (see Figure 4.3) instead is more difficult to interpret and doesn't show such a high correlation as the the total mutual information but we can see that in coincidence with certain peaks in the STLFSI also the total transfer entropy of the network peaks.

While the system aggregated measures give us information regarding the overall system financial stress through bank level measures we would like to investigated which banks helps to predict the financial stress indexes. The bank level timeseries are computed aggregating the single edges mutual information and transfer entropy according to the procedure illustrated in Section 4.3.3 for each market day. After these pre-processing we obtain three timeseries for each single bank: mutual information (the total mutual information between the bank and its neighbours), transfer entropy inflow (the sum of the transfer entropy incoming to the bank from its 1-day lagged neighbours), transfer entropy outflow (the sum of the transfer entropy going from the 1-day lagged bank node to its contemporary neighbours). These three measures summarize, respectively, three different types of information; i) how much a bank returns are correlated with the rest of the banking system (more precisely with its neighbours); ii) how much knowing a bank returns helps predicting the rest of the system returns; iii) how much knowing the system returns helps to predict a bank returns. We want to investigate whether some of these bank level measures helps to predict financial stress indexes

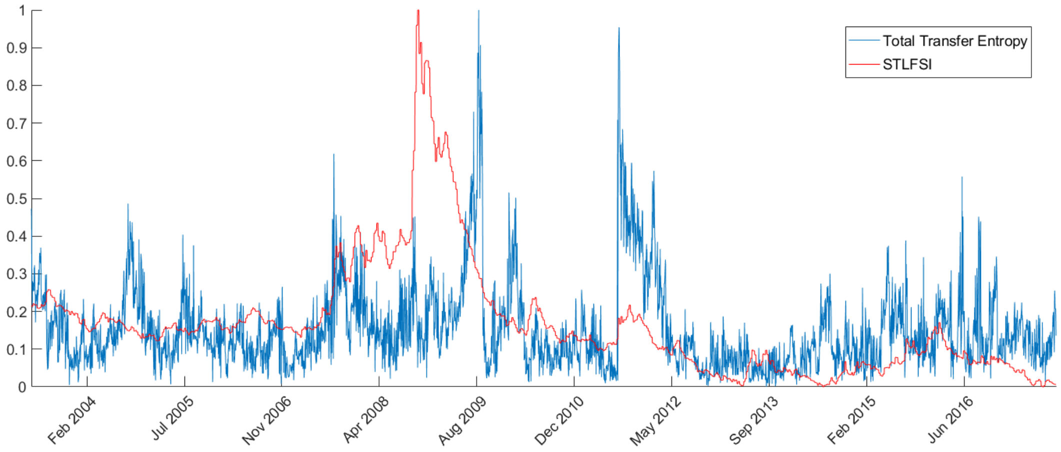


FIGURE 4.3: Total network transfer entropy vs STLFSI trends comparison

like STLFSI, Libor-OIS spread and USD/CHF exchange rate. From such a finding we could understand that a bank has an important role in the financial stress dynamics of the system. To test if these measures help to predict the financial stress indexes we resort to the Granger causality test and in particular a recent extension of it, the partial Granger causality [Guo et al. 2008]. We resort to the partial Granger causality to mitigate the possible confounding influence in the eventuality of missing and latent variables [Pearl 1999] as explained in Section 4.3.5. When testing for causality we condition on three macroeconomic variables to control the effect of the macroeconomic cycle and eventual spurious correlations. These three control variables are the US 10Y Treasuries yield, the gold price and the EUR/USD exchange rate and are related to the general economic conjuncture. For the partial Granger causality test we resort to the R package FIAR (Functional Integration Analysis in R) [Roelstraete et al. 2011]. We test the linear partial Granger causality from the bank level timeseries to the different financial stress indexes for three different periods in which we split the analysis: pre-crisis (2003-2006), financial crisis (2007-2010), post-crisis (2011-2017). Prior to testing for causality, the timeseries have been tested for stationarity with a Dickey-Fuller test and where necessary the time series have been differentiated with the "forecast" package available in R [Hyndman et al. 2008, Hyndman et al. 2018]. Each causality test is performed considering up to the 5^{th} lag for the bank level time series. In Table 4.2 we report the results of the partial Granger causality test.

Analyzing the results in Table 4.2, where we list the banks with at least two significant lags at $\alpha = 0.05$, we can see that the statistically significant banks comprehend both large banks like JPM, C, WFC, medium size banks like STL, ASB and smaller banks like NYCB. The list includes also large foreign banks like BBVA, HSBC and DB that have considerable activities in the US.

In Table 4.3 we report the banks that appear more frequently among the statistically significant banks within each period and for each bank level measure. Thus, given all the banks that have at least 2 significant lags at $\alpha = 0.05$ within a period and one bank level measure we select by majority voting those that appear most frequently. The logic behind this choice is that if a bank is very relevant in predicting the stress indexes, it should give causality signals to all the stress indicators (STLFSI, Libor-OIS spread and CHF/USD rate returns). So, the most significant banks should have significant lags in predicting not only one stress index but possibly more. In this table we see many of the largest US banks, in particular when the transfer entropy outflow is considered. This is reasonable because largest banks are more likely to influence the rest of the system and thus their stock returns should help in predicting the returns of the other banks.

Period	Bank level measure	Financial stress index	Statistically significant banks
'03-'06	Mutual Information	CHF/USD returns	ASB, CBSH, PACW, TCF, UMPQ
'03-'06	Mutual Information	Libor-OIS spread	BBT, PNC
'03-'06	Mutual Information	STLFSI	AF, UMPQ
'03-'06	Transfer Entropy out.	CHF/USD returns	BANC, IBKC, OZRK
'03-'06	Transfer Entropy out.	Libor-OIS spread	AF, BANC, PBCT
'03-'06	Transfer Entropy out.	STLFSI	BOKF, FNB, SAN, SCHW, SIVB
'03-'06	Transfer Entropy in.	CHF/USD returns	FBP, IBKC, NYCB, RF
'03-'06	Transfer Entropy in.	Libor-OIS spread	FCNCA, FITB, PACW, SIVB
'03-'06	Transfer Entropy in.	STLFSI	BMO, FITB, FULT
'07-'10	Mutual Information	CHF/USD returns	AXP, BK, DB, SBIN, SIVB, UCBI, WBS
'07-'10	Mutual Information	Libor-OIS spread	CMA, DB, FHN, FITB, HBAN, SAN, TD, UCBI
'07-'10	Mutual Information	STLFSI	BANC, BK, PNC, STL
'07-'10	Transfer Entropy out.	CHF/USD returns	GS, MTB, SLM
'07-'10	Transfer Entropy out.	Libor-OIS spread	ASB, BAC, BBT, BBVA, COF, JPM, TCF
'07-'10	Transfer Entropy out.	STLFSI	BAC, BANC, BCS, COF, JPM, STI
'07-'10	Transfer Entropy in.	CHF/USD returns	BBT, BXS, KEY, SBIN, WFC
'07-'10	Transfer Entropy in.	Libor-OIS spread	HBAN, HSBC, PBCT, SBIN, WBS
'07-'10	Transfer Entropy in.	STLFSI	AF, BANC, SBIN, WBS
'11-'17	Mutual Information	CHF/USD returns	BOKF, C, CFR, FBC, JPM, MTU, RF, RJF, STL
'11-'17	Mutual Information	Libor-OIS spread	BANC, C, CFR, RJF, STL
'11-'17	Mutual Information	STLFSI	C, CFR, RJF, SAN, SLM, STI, STL, VLY
'11-'17	Transfer Entropy out.	CHF/USD returns	
'11-'17	Transfer Entropy out.	Libor-OIS spread	ASB, BOKF
'11-'17	Transfer Entropy out.	STLFSI	NYCB, SF, UMBF
'11-'17	Transfer Entropy in.	CHF/USD returns	
'11-'17	Transfer Entropy in.	Libor-OIS spread	BBVA, TD, WFC
'11-'17	Transfer Entropy in.	STLFSI	CFR, OZRK, WFC

TABLE 4.2: Partial Granger causality results: causality test from bank level measures to financial stress indexes; statistically significant banks have at least two significant lags at $\alpha = 0.05$

Period	Bank level measure	Most significant banks
2003-2006	Mutual Information	
2003-2006	Transfer Entropy outflow	BANC
2003-2006	Transfer Entropy inflow	FITB
2007-2010	Mutual Information	BK, DB
2007-2010	Transfer Entropy outflow	BAC, COF, JPM
2007-2010	Transfer Entropy inflow	SBIN
2011-2017	Mutual Information	C, CFR, RJF, STL
2011-2017	Transfer Entropy outflow	
2011-2017	Transfer Entropy inflow	WFC

TABLE 4.3: Most significant banks: list of the banks that appear more times in the Statistically significant banks within a period and fixed the bank level measure in Table 4.2

In the pre-crisis period '03-'06 we find less significant banks, specially when testing the partial Granger causality for the STLFSI and Libor-OIS spread. This is expected since both the indexes have been widely adopted and regarded during and after the crisis. In particular, STLFSI has been developed after the crisis and backward calculated with the goal of being a good indicator for the crisis. Moreover both the indicators during the pre-crisis period were not subject to sudden and extensive spikes or changes thus is more difficult that a single bank stock is useful in predicting its behaviour.

During the crisis period '07-'10 there there are more banks with statistically significant p-values in the partial Granger causality tests due to the greater correlation of the entire system, found also by other studies [Dilip et al. 2013]. This is also in agreement with the fact that both the total mutual information and the transfer entropy of the network peak during the crisis. During the crisis is interesting to look at the banks whose transfer entropy outflow is most relevant in Granger causing the indexes. These banks in fact, are those whose influence on the rest of the system (transfer entropy outflow) is more useful to predict the stress indexes; they are mainly large banks (Bank of America, Capital One Financial, J.P. Morgan) that had an important role during the crisis.

Bank of America (BAC) is the second largest financial institution in the US and has been severely affected by the crisis. Several acquisitions in fact, had increased its exposition towards consumer credit and house mortgages. In 2005 it bought the credit card giant MBNA, in 2008 it acquired Countrywide Financial, the largest mortgage originator in America at the time and the troubled stockbroker Merrill Lynch. All of these businesses registered enormous losses during the crisis.

Capital One Financial (COF) in mid 2007 in fact, announced that it would have eliminate 1,900 jobs and shut down a wholesale mortgage unit it had acquired less than a year before, in response to the U.S. housing downturn, and posted great losses.

J.P. Morgan (JPM) has been a protagonist bank during the crisis in positive terms compared to the others. J.P. Morgan in fact, in the years prior to the crisis mostly avoided subprime mortgages, structured investment vehicles and collateralized debt obligations. When the subprime bubble triggered a massive deleveraging J.P. Morgan was mostly unharmed compared to its rivals. So J.P. Morgan was in such a good position, that it offered to take over Bear Sterns.

During the post-crisis period, '11-'17 we register less statistically significant banks, in line with the intervention of the central banks whose policies have helped cooling down the financial system. Among the most relevant, we find both large banks like Wells Fargo (WFC) and Citigroup (C) and smaller institutes like Frost bank (CFR), Raymond James bank (RJF) and Sterling National bank (STL). The two large banks are bad performers among their peers. Wells Fargo while recovering from the crisis has witnessed a troubled post-crisis period studded with lawsuits and scandals that have undermined its reputation at the point that in 2018 the bank launched a marketing campaign called "Re-Established" to emphasize the company's commitment to re-establish trust with stakeholders. Citigroup after the government bailout, has failed FED stress test in 2012 and 2014 and has seen a period of downsizing characterized by market exits, sell-off and shutdowns of different units. Instead the smaller statistically significant institutes (CFR, RJF and STL) are all characterized by an intense expansion and acquisition activity during the post crisis period.

It is also important to note that there are certain banks that are significant in more than one time period. For example, ASB, BANC and SAN are significant in all the three periods ('03-'17), while SIVB, TCF, BBT, PNC are significant before and during the financial crisis ('03-'10) and STL, SLM, JPM and STI ('07-'17) are significant during and after the crisis. These banks comprehend both important hubs in the network model like ASB, SIVB and STL and more peripheral nodes like TCF or SAN (see Fig. 4.1)

4.6 Conclusions

In this work we have presented two main contributions. Firstly, we have applied a recently presented graphical model inference methodology LoGo in the investigation of U.S. Banks stock returns to understand the network structure and its evolution from 2003 to 2017. Thanks to the LoGo computational efficiency we could estimate a separate graphical model for each market day and generate several time series of bank related measures computed from the network structure. Secondly, we have presented a way to leverage the graphical model information comparing the measures derived from its structure with well known financial stress indexes and performed a causality analysis among them. To perform the causality analysis we resorted to the partial Granger causality method to take in consideration different control variables.

The inferred graphical models and the bank related measures extracted from them have shown to be an interesting tool for monitoring the U.S. bank system evolution. The bank related measures extracted from the network in fact, have shown correlation with several financial stress indexes and to be linked in Granger causality terms to some of them acting as causing variables in the different time frames.

Considering further research on this topic, it would be interesting to use other publicly available information on banks as well, like for example, bonds issued by banks or banks CDS. Bonds and CDS may capture different risk information more related to the bank default risk. In this case it would be necessary to handle different maturities in a proper way in order to obtain comparable variables. Finally it would be possible to merge the different networks obtained in a multilayer network model that could potentially capture different aspects of the bank risk.

Chapter 5

Assessing news contagion in finance

5.1 Summary

The analysis of news in the financial context has gained a prominent interest in the last years. This is because of the possible predictive power of such content especially in terms of topics, topics proportions and associated sentiment. In this chapter, we focus on a specific aspect of financial news analysis: how the covered topics modify according to country and time dimensions. To this purpose, we apply a modified version of the LDA Topic Model, the so-called Structural Topic Model (STM), that takes into account numerical and categorical covariates as well. Our aim is to study the possible evolution of topics extracted from two well known news archive (Reuters and Bloomberg) and to investigate a causal effect in the diffusion of the news by means of a Granger causality test. Our results show that both the temporal dynamics and the spatial differentiation matter in the news contagion.

5.2 Introduction

With the rapid growth of online information, text analysis and categorization have become core topics in many different disciplines ranging from politics to finance and social sciences in general. Text analytics techniques are an essential part of text mining and are used to classify documents (of any kind) and to find interesting information therein.

The interpretation of text by machines, the task of natural language processing (NLP), is complex due to the richness of human language, as well as the ambiguity present at many levels, including the syntactic and semantic ones. From a computational point of view, processing language means dealing with sequential, highly variable and sparse symbolic data, with surface forms that cover the deeper structures of meaning. Despite these difficulties, there are several methods able to extract part of the information content present in collections of texts. Some of these rely on handcrafted features, while others that are data driven exploit statistical regularities in language and often rely on word representations. Class based models, for example, learn classes of similar words based on distributional information, such as Brown clustering [Brown et al. 1992] and Exchange clustering [Martin et al. 1998, Clark 2003]. Soft clustering methods, such as Latent Semantic Analysis (LSA) [Landauer et al. 1998] and Latent Dirichlet Allocation [Blei et al. 2003], associate words to topics through a distribution over words of how likely each word is in each cluster/topic. In the last years, many contributions employ neural networks and semantic vector representations [Hochreiter and Schmidhuber 1997, Mikolov et al. 2013, Pennignton et al. 2014, Cho et al. 2014] to model complex and non-local relationships in the sequential input [Socher et al. 2011, Socher et al. 2013, Collobert et al. 2011, Kalchbrenner et al. 2014]. If we focus specifically on the finance related research area, we can list several papers that take advantage of text analytics per se or as an additional source of information to be used. Central banks themselves have been recently starting to recognize the utility of text data in financial risk analytics [Bholat et al. 2015, Hokkanen et al. 2015].

In this chapter, we follow a stream of research based on official news and we deepen a particular aspect: improving information elicitation to enhance the model with contextual information (metadata and covariates) related to the characteristics and environment in which the entities of interest are operating to discover and analyze contagion patterns in the information flow. Indeed, the introduction of contextual information in the models is not a straightforward process but requires a careful choice of the additional information provided in order not to introduce additional noise. The addition of metadata aims at increasing the potential value of text as a source in data analysis [Soo 2013]. More in detail, we choose as covariates temporal and spatial variables, so to help the understanding of possible evolution pattern or contagion effects in the information flows. In this respect, we employ a modified version of the well-known Latent Dirichlet Allocation topic model called Structural Topic Model (STM) proposed by [Roberts et al. 2016] that explicitly includes covariates in the model fitting. To our knowledge, this is the first attempt to assess the contagion effect through news in finance. In particular, we propose to analyze banks' related news and correlate the news topics with the banks' nationality and the news time stamp aggregated at either monthly and weekly basis.

This recent rise of interest around the integration of text-based computational methods for the assessment of financial risk is fuelling a rapidly growing literature that can be divided in two main streams according to the type of textual source: social media blogs and platform (namely Twitter, Facebook, and Google Trends) or official news archive (above all, Reuters and Blomberg).

In the first case, the constant production of detailed online information streaming from social networking and micro-blogging platforms, is increasingly attracting the attention of researchers and practitioners especially for the detection and monitoring of sentiments and opinions. Indeed, social media contents may constitute a relevant asset for financial institutions to gain useful insights about the clients' needs and perceptions in real time. Insofar, extracting sentiments from Twitter has been already employed for several purposes: to predict the trends of Dow Jones Index [Bollen et al. 2011]; to check the effects of sentiments on stock price and volume in the Dow Jones Index [Ranco et al. 2015]; to predict market prices in the Italian financial market [Cerchiello and Giudici 2015]; or to estimate Italian banks systemic risk like in Chapter 3. There are many other papers in this field leveraging Twitter for financial analysis and prediction [Sprenger and Welpe 2010, Brown 2012, Mittal and Goel 2012, Rao and Srivastava 2012, Nann et al. 2013, Oliveira et al. 2013]. Another strand of literature uses social media as an alternative way to release information, thus reducing information asymmetry and improving stock liquidity, attracting more investors. Other papers, such as [Chawla et al. 2016] or [Giannini et al. 2013], use Twitter data dynamically to investigate how information diffusion affects trading and how tracks changes in investor disagreement.

On the other hand, if we consider official news as source of information, not only sentiment but also content analysis is crucial, since the resulting outcomes are used for assessing correlation with events of interest (typically stress events). Many of the proposed approaches have been based on hand-crafted dictionaries that, despite requiring work to be adapted to single tasks, can guarantee good results due to the direct link to human emotions and the capability of generalizing well through different datasets [Nyman et al. 2015, Soo 2013]. The first analyzes sentiment trends in news narratives in terms of excitement/anxiety and find increased consensus to reflect pre-crisis market exuberance, while the second correlates the sentiment in news with the housing market. Despite the good results, there are applications where it could be preferable to avoid dictionaries in favour of more data driven methods, which have the advantage of higher data coverage and capability of going beyond single word sentiment expression [Malo et al. 2014]. provide an example of a more sophisticated supervised corpus-based approach, in which they apply a framework modelling financial sentiment expressions by a custom dataset of annotated phrases. In the last years, different papers, embracing the data driven approach, have used the deep learning models to analyze textual data. They have shown good results in predicting distress events of financial institutions like [Rönnqvist and Sarlin 2017] and the following Chapter 7 and in predicting S&P500 stocks [Ding et al. 2015].

This work continues pursuing this line of research by applying a fully unsupervised data driven model based on topic modelling, supervised only by a posterior interpretation of the discovered topics.

The rest of the chapter is organized as follows: in Section 5.3, we illustrate the applied model; in Section 5.4, we describe the data and the preprocessing steps; in Section 5.5, the results are presented; in Section 5.6, conclusions of the work with hints on future developments are discussed.

5.3 Methodology

Text analysis is a complex task that poses several different issues ranging from the problem of polysems (multiple senses for given words) and synonyms (same meaning for different words) to the computational effort and allocation of largely sparse data matrices. One of the first effective models able to solve some of those issues is represented by Latent Semantic Analysis (LSA) [Deerwester et al. 1990]. The basic idea of LSA is to work at semantic level by reducing the vector space through Singular Value Decomposition (SVD), producing occurrence tables that are not sparse and that help in discovering associations between documents. To establish a solid theoretical statistical framework in this context, [Hofmann 1999] proposed a probabilistic version of LSA (pLSA). Such model, also known as the aspect model, is rooted in the family of latent class models and is based on a mixture of conditionally independent multinomial distributions for modeling the words-documents pair. The intention from the introduction of pLSA was to offer a formal statistical framework, helping the parameter interpretation issue as well. The goal was achieved only partially, because the multinomial mixtures, whose components can be interpreted as topics, offer a probabilistic justification at words but not at documents level. In fact, the latter are represented merely as a list of mixing proportions derived from mixture components. Moreover, the multinomial distribution presents as many values as there are in the training documents and therefore it learns topic mixture on those trained documents. The extension to previously unseen documents is not appropriate since there can be new topics. To overcome the asymmetry between words and documents and to provide a fully generative model, [Blei et al. 2003] proposed the LDA (Latent Dirichlet Allocation) model. LDA is still based on the “bag of words” assumption that neglects the word order in the text is a fully generative model since it posits a Dirichlet distribution over documents in the corpus, while each topic is drawn from a Multinomial distribution over words. However, note that [Girolami and Kaban 2003] have shown that LDA and pLSA are equivalent if the latter is under a uniform Dirichlet prior distribution. LDA does not solve all the challenges of involved in topic modelling and the main restriction embedded in its approach (due to the Dirichlet distribution) refers to the assumption of independence among topics. To tackle this issue Correlated Topic Model (CTM), have been proposed in [Blei and Lafferty 2006]. CTMs introduce correlations among topics by replacing the Dirichlet random variable with the logistic normal distribution. Unlike LDA, CTM presents a clear complication in terms of inference and parameter estimation since the logistic normal distribution and the Multinomial are not conjugate. To bypass the problem, the most recent alternative is represented by the Independent Factor Topic Models (IFTM) introduced in [Putthividhya et al. 2009]. Such proposal makes use of a latent variable model approach to detect hidden correlations among topics. The choice to explore the latent model world allows to choose among several alternatives ranging from the type of relation, linear or not linear, to the type of prior to be specified for the latent source.

In this chapter, we focus on one of the most recent extensions of the LDA model proposed by [Roberts et al. 2016]. This model, called Structural Topic Model (STM), considers the explicit inclusion of covariates that can help in describing and interpreting the topics along the corpus. More specifically, STM allows for covariates to influence two elements of the model: the topic prevalence and the topical content. With the former, the authors refer to the proportion of a document devoted to a topic, while the latter describes the word rates used in discussing a topic. The authors take advantage of the Generalized Linear Models framework to accommodate for general covariate information (or metadata) into topics model thanks also to two previous papers from [Mimno and McCallum 2008] and [Eisenstein et al. 2011].

Since STM depends upon LDA, we first summarize the latter and then we move to the former. [Blei et al. 2003] defines the model as follows:

$$\theta_i \sim Dir(\alpha) \tag{5.1}$$

$$\phi_k \sim Dir(\beta) \tag{5.2}$$

$$z_{ij}|(\theta_i) \sim Multinomial(\theta_i) \tag{5.3}$$

$$x_{ij}|z_{ij} \sim Multinomial(\phi_{z_{ij}}) \tag{5.4}$$

where d_i for $i = 1, \dots, N$ is collection of N documents and words $\{x_{ij}\}_{j=1}^{J_i}$ within each document d_i listed in a common vocabulary containing V words, with N the number of documents and J_i the number of words in the document d_i. Assuming that we have k topics for $k = 1, \dots, K$, θ_i is the length-K per document topic distribution for document d_i, ϕ_k is the length-V per topic word distribution for the k-th topic and z_{ij} is the topic for the j-th word in d_j. Finally, α and β are hyperparameters that influences respectively the documents distributions over topics and the topics distributions over words.

Coming to the Structural Topic Model, [Roberts et al. 2016] defines it as follows:

$$\theta_i|(C_i\gamma, \Sigma) \sim LogisticNorm(C_i\gamma, \Sigma) \tag{5.5}$$

$$\phi_{ik} \propto exp(m + k_k + k_{c_i} + k_{k_{c_i}}) \tag{5.6}$$

$$z_{ij}|(\theta_i) \sim Multinomial(\theta_i) \tag{5.7}$$

$$x_{ij}|z_{ij} \sim Multinomial(\phi_{iz_{ij}}) \tag{5.8}$$

where $w = 1, \dots, W$, $k = 1, \dots, K$, C_i is the covariates matrix, γ is the coefficient vector, Σ is the covariance matrix, ϕ_{ik} is the word distribution for document d_i and k-th topic, m is a reference log-word distribution while k_k, k_{g_i} and $k_{k_{g_i}}$ represent the deviations from the baseline due, respectively, to the topics, the covariates and their interaction effect.

The strength of the model relies on its three different components clearly represented in Equations 5.5-5.8: the topic prevalence is modelled by Equation 5.5 through a logistic normal distribution which mean is not constant but it depends on the covariates. The topical content is represented by Equation 5.6 according to which the word occurrences are modelled in terms of log-transformed rate deviations from a corpus based distribution m. The parameters k_k, k_{g_i}, $k_{k_{g_i}}$ represent the specific deviations: respectively for the topic, for the covariates and for the interaction topic-covariates. Finally, Equations 5.7 and 5.8 comprise the central part of the model reporting the distribution of topics z_{ij} and of words x_{ij} both sampled from a Multinomial distributions. LDA and STM are similar in the core language of the model that is the sampling mechanism of the topics and of the words as appear from Equations 5.3, 5.4, 5.7 and 5.8. The main difference is in the parameters of the Multinomials that, for the STM model, depend upon covariates.

Since our research hypothesis aims at analyzing a contagion effect and its patterns in the diffusion of topics among countries according to a temporal dimension, we need a method to assess such effect. In the following paragraph, we introduce the Granger causality test, a well-known econometric test useful when causality is the object of interest.

Granger causality entails the statistical notion of causality based on the relative forecast power of two time series. Time series j is said to "Granger-cause" time series i if past values of j contain information that helps in predicting i above and beyond the information contained in past values of i alone.

In a well known paper [Granger 1969], Granger has proposed a useful test based on the following principle: if lagged values of time series X_t contribute to foresee current values of time series Y_t in a forecast achieved with lagged values of both X_t and Y_t, then we say X_t *Granger causes* Y_t. As was first shown in [Sims 1972], the Granger causality corresponds to the concept of exogeneity and it is therefore necessary to have a unidirectional causality in order to guarantee consistent estimation of distributed lag models. The mathematical formulation of this test is based on linear regressions of X_{t+1} on X_t and Y_t

In our research framework, we propose to calculate the Granger causality test on pairs of times series defined as follows:

- R_t^{kq}: given a topic k, the vector of document counts showing a topic prevalence θ_i^k larger than a specified threshold with regards to country q at time t.
- R_t^{kp}: given a topic k, the vector of document counts showing a topic prevalence θ_i^k larger than a specified threshold with regards to country p at time t.

To ease the notation, we refer to R_t^q given the topic k and country q (similar to country p).

Thus, applying the test for a given pair of count vectors for topic k and countries $q-p$, we result in fitting the following equations:

$$R_{t+i}^q = \beta_0^q R_t^q + \beta_1^{qp} R_t^p + e_{t+i}^q \tag{5.9}$$

$$R_{t+i}^p = \beta_0^p R_t^p + \beta_1^{pq} R_t^q + e_{t+i}^p \tag{5.10}$$

Our null hypothesis is therefore: $H_0 : \beta_1^{qp} = \beta_1^{pq} = 0$. Taking into account that we are dealing with monthly time series and weekly time series, in our tests, we have considered up to two lags as plausible windows of analysis.

5.4 Data

The data analyzed are contained in two public financial news dataset extracted by Reuters News and Bloomberg News containing respectively 106,521 and 447,145 documents[1]. The data span a period from October 2006 to November 2013. Such time frame is very interesting from a financial perspective since it comprehends the sub-prime crisis started in 2007 and its following evolution with modest recovery and the beginning of the sovereign debt crisis. Moreover, beside these major topics, there have been many spot hot topics which have

[1] The datasets are available on the Github of Philippe Remy at https://github.com/philipperemy/financial-news-dataset and have been retrieved and appropriately collected using Python.

periodically grabbed the attention of the media like, for example, the Madoff fraud, Barclays and Deutsche bank Libor manipulation investigation and UBS tax evasion controversy.

The datasets contain a broad variety of articles ranging from analysts' recommendations to earning announcements to legal investigation news. All the news report the timestamp of the corresponding day. The datasets need to be carefully inspected and cleaned up according to the purpose of the analysis. In our case, the analysis focuses on the SIFIs banks (Systemically Important Financial Institution according to Basel Committee definition) listed in Table 5.1 and thus we cleaned the dataset to reduce as much as possible the non-bank related news. Then, we have tokenized each document into sentences and kept only those containing SIFI labels (see Table 5.1). We have developed a dictionary of bank names to be matched with the available sentences and we do not include bank tags and tickers due to their possible ambiguity with other entities (for example City Group ticker C and Santander SAN can easily refer to other non-related arguments). In addition, to associate a phrase to a single bank and to avoid multiple imputation, we have kept sentences referring only to one bank. Finally, since many of these institutions are very active in the investment banking sector and often release reports on other companies, we have dropped the sentences containing keywords associated with this kind of news, such as "analyst", "analysts", "said", "note", "report", and "rating". These words have been easily detected by looking at the wordclouds referred to such news. This selection procedure is somehow restrictive, but it is necessary to deal with a clean dataset focused only on banks related news. The phrases remaining after this filtering are 136,419 and cover many of the SIFI with the proportions reported in Table 5.1.

Bank	# of Sentences	Country
Bank of America	19,203	USA
Goldman Sachs	16,258	USA
Citigroup	15,446	USA
UBS	13,414	Switzerland
Barclays	11,434	UK
Morgan Stanley	11,162	USA
HSBC	8,693	UK
Deutsche Bank	7,471	Germany
Credit Suisse	6,385	Switzerland
Wells Fargo	4,876	USA
Bank of China	3,416	China
Societe Generale	2,463	France
BNP Paribas	2,012	France
Royal Bank of Scotland	1,943	UK
Standard Chartered	1,813	UK
Commerzbank	1,512	Germany
BNY Mellon	1,427	USA
Credit Agricole	1,195	France
Banco Santander	1,023	Spain
State Street	926	USA
Sumitomo Mitsui	900	Japan
JP Morgan	755	USA
Industrial and Commercial Bank of China	732	China
BBVA	718	Spain
Lloyds Bank	648	UK
China Construction Bank	387	China
ING Bank	110	Netherlands
Unicredit	94	Italy
Dexia Group	2	Belgium
Total	136,418	

TABLE 5.1: List of considered SIFI Banks.

In Table 5.2, we report the number of sentences grouped by country. It clearly appears that the distribution of the sentences across the country is heterogeneous and this has an impact on the comparability of results across banks and countries. Thus countries, not showing enough news have been excluded from the analysis according to criteria we explain below.

To fit the STM model, we need to choose appropriate covariates that we consider relevant in the description of the topics. To this purpose, we have considered a temporal variable reporting the month or the week in which the news have been released. For sake of comparability and robustness, the analysis has been carried out with two different versions of the Reuters-Bloomberg dataset with regards to the temporal dimension. Thus, the time covariate has been considered according to two different aggregation periods:

- Monthly-based: The time stamp of each news has been grouped on a monthly basis, obtaining 85 months starting with October 2006 (Month 1) and ending with November 2013 (Month 85).

- Weekly-based: The time stamp of each news has been grouped on a weekly basis, obtaining 370 weeks starting with 23^{rd} October 2006 (Week 1) and ending with 19th November 2013 (Week 370).

This allowed us to fit and compare two different configurations of the STM: the first one assuming a monthly contagion effect, the second one, indeed more realistic, a weekly contagion transmission. However, aggregating the news at week level has a important impact on the list of SIFI banks that can be reliably evaluated: we do not have enough news on a weekly basis for all banks, thus we must consider only the ones most covered by the media. As a result, in the monthly based analysis, we include with 25 banks, that are those having at least 10 mentioning sentences per month on average or at least 1,000 mentioning sentences during the considered period. In the weekly based analysis we consider 10 banks, that are those having at least 10 mentioning sentences per week on average during the considered period.

Country	# of Sentences
USA	70,053
UK	24,531
Switzerland	19,799
Germany	8,983
France	5,670
China	4,535
Spain	1,741
Japan	900
Netherlands	110
Italy	94
Belgium	2
Total	136,418

TABLE 5.2: Distribution of documents per country.

Along with the temporal variable, we have considered a spatial information mapping each SIFI banks onto the corresponding country (namely the country in which the headquarters is based). Then, we have introduced as many dummy variables as the involved countries: five in the monthly based case (France, Germany, Switzerland, UK and USA) and four in the weekly based case (Germany, Switzerland, UK and USA).

The rationale behind the inclusion of temporal and spatial covariates is the following: while the formers help us in monitoring the evolution of news along the time horizon, the latter is useful in disentangling the country/institution effect.

5.5 Results

To select a model with a good interpretability, we have tested different topic numbers and manually inspected the results. To evaluate the clarity of the resulting topics, we have considered the top 20 words associated to each topic according to the highest probability measure and to the frequency measure "FREX". In [Roberts et al. 2016b], the FREX metric has been proposed to measure exclusivity in a way that balances word frequency. The FREX is the weighted harmonic mean of the word's rank in terms of exclusivity and frequency within the topic.

We tested six different configurations for the monthly based analysis with 5, 10, 12, 15, 25, and 35 topics (simulation time in Table 5.3), and we concluded that results with 10, 12 and 15 topics are stable and consistent with each other in terms of identified arguments (see Table 5.4). We also tested different configurations for the weekly based analysis with 10, 15 and 25 topics (with simulation times analogous to the monthly case), and we concluded that results with 15 topics are consistent with the monthly case with 15 topics in terms of identified arguments.

# of Topics	**Time (s)**
5	371
10	522
12	685
15	543
25	1,155
35	6,667

TABLE 5.3: Simulation time of the different STM configurations.

	Monthly Aggregation			**Weekly Aggregation**
Topic Title	**10 Topics**	**12 Topics**	**15 Topics**	**15 Topics**
UBS tax fraud scandal	Y	Y	Y	Y
Market performance	Y	Y	Y	Y
Stock recommendation	Y	Y	Y	Y
Chinese companies news	Y	Y	Y	-
Hedge Funds, Private Equity and Inv. Banking	Y	Y	Y	Y
Press comments and PR	Y	Y	Y	Y
Citigroup bailout	Y	Y	Y	Y
Advisory	-	-	Y	-
Morgan Stanley Investment Banking	Y	Y	Y	Y
Euro area banks	Y	Y	Y	-
Madoff scandal	-	-	Y	Y
Barclays and Deutsche B. LIBOR manipulation	Y	Y	Y	Y
Bond, Equity,and CDS markets	-	-	Y	Y
Mortgage crisis	-	Y	Y	Y
Spanish banks	-	-	Y	-
General view on the economy	-	Y	-	-
Insider trading investigation	-	-	-	Y
Wells Fargo-Wachovia acquisition	-	-	-	Y
Bank management changes	-	-	-	Y
US banks stocks performance	-	-	-	Y

TABLE 5.4: STM configurations comparison on monthly and weekly aggregated data.

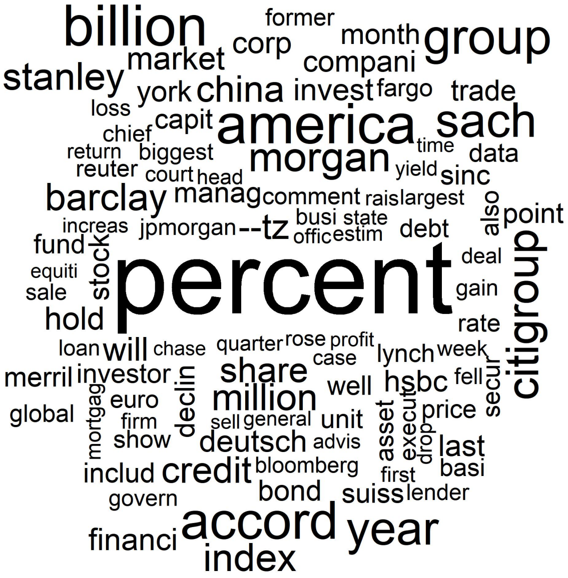
For comparability and reproducibility, in each simulation run, we applied the same data cleaning process removing English stopwords, keeping only the words with length between 4 and 15 letters appearing in more than 30 and less than 45,000 documents. We kept also the STM model parameter in R set to an Expectation Maximization improvement tolerance equal to 1×10^{-5} (as suggested by the package developers and by empirical evidence). In the following paragraphs, we describe the 15 topics model configuration since it shows well defined and interpretable topics. Moreover, as emerges from Table 5.4, it is fully comparable to other configurations such as 10 or 12 topics, but with an increased level of clarity and definition and with the addition of relevant topics such as "Madoff scandal" and "Spanish banks news".

Our findings show that the identified topics represent some of the most discussed financial events that took place between 2007 and 2013, in particular:

"UBS tax fraud" (Topic 1), "Market performance" (Topic 2), "Stock recommendation" (Topic 3), "Chinese companies news" (Topic 4), "Hedge Funds, Private Equity and Investment Banking" (Topic 5), "Press comments and PR" (Topic 6), "Citigroup bailout" (Topic 7), "Advisory" (Topic 8), "Morgan Stanley Investment Banking" (Topic 9), "Euro area banks" (Topic 10), "Madoff fraud scandal" (Topic 11), "Barclays and Deutsche Bank LIBOR manipulation" (Topic 12), "Bond, Equity and CDS markets" (Topic 13), "Mortgage crisis" (Topic 14), and "Spanish banks" (Topic 15). For completeness, we report in Table 5.5 the complete list of words associated to each topic according to the FREX measure that accounts for both their overall frequency and exclusivity to the specific topic.

The wordcloud in Figure 5.1 reports the most relevant words along the whole analyzed corpus and it clearly highlights some words specifically connected to the 15 topics such as Citigroup, Barclays, Morgan, mortgage, etc.

FIGURE 5.1: Wordcloud of the 15 topics analysis.



Topic	Words
Topic 1	FREX: charg, justic, guilti, account, ubsn, evas, plead, prosecut, crimin, hide, depart, evad, client, indict, california, avoid, wealthi, adoboli, involv, ubsnvx
Topic 2	FREX: gain, percent, cent, cmci, lost, ralli, advanc, drop, materi, sinc, jump, return, slip, tumbl, climb, slid, compil, rose, close, bloomberg
Topic 3	FREX: sumitomo, mitsui, suiss, csgn, scotland, neutral, credit, lloy, spectron, neutral, rbsl, royal, icap, mizuho, csgnvx, maker, suisse , outperform, baer
Topic 4	FREX: elec, cosco, sino, comm, lung, chem, pharm, fook, sang, shougang, yuexiu, sinotran, picc, swire, people , intl, emperor, shui, citic, hang
Topic 5	FREX: sach, goldman, groupinc, blankfein, sachs , gupta, rajaratnam, sachsgroup, corzin, paulson, vice, wall, rajat, tourr, presid, warren, buffett, obama, hathaway, gambl
Topic 6	FREX: spokesman, comment, charlott, spokeswoman, immedi, carolina-bas, tocom, bacn, countrywid, north, avail, lewi, moynihan, confirm, carolina, declin, respond, corp, repres, america
Topic 7	FREX: bailout, citigroup, pandit, sharehold, prefer, receiv, vikram, troubl, citigroup, announc, rescu, common, taxpay, worth, subprim, crisi, dividend, loss, plan, shed
Topic 8	FREX: advis, hire, head, team, familiar, privat, wealth, manag, appoint, deal, equiti, arrang, advisori, co-head, counsel, person, barclay, financ, dbkgnde, advic
Topic 9	FREX: stanley, morgan, stanley , smith, barney, gorman, mack, ventur, facebook, estat, bear, fuel, brokerag, underwrit, real, stearn, crude, commod, brent, healthcar
Topic 10	FREX: societ, pariba, commerzbank, euro, estim, profit, quarter, french, general, forecast, itali, greek, half, predict, germany , technic, germani, greec, socgen, incom
Topic 11	FREX: case, mellon, truste, southern, district, york, suit, bankruptci, mortgage-back, claim, stempel, oblig, collater, file, madoff, lehman, picard, jonathan, rakoff, manhattan
Topic 12	FREX: libor, manipul, diamond, regul, scandal, told, wrote, think, confer, fine, ubss, gruebel, respons, lawmak, event, england, polici, hsbcs, complianc
Topic 13	FREX: basi, point, markit, itraxx, percentag, yield, basispoint, swap, spread, preliminari, manufactur, extra, read, managers , tokyo, demand, releas, bond, econom, narrow
Topic 14	FREX: fargo, charter, chase, well, standard, jpmorgan, jpmn, home, wfcn, build, korea, portfolio, loan, francisco-bas, origin, size, mutual, small, fargo , india
Topic 15	FREX: banco, santand, bbva, bilbao, peso, spain , argentaria, spanish, chile, vizcaya, brazil, latin, mexico, spain, brasil, follow, mover, brazilian, mexican

TABLE 5.5: List of 15 topics obtained from monthly aggregated data. The associated words are ordered by FREX measure (words are weighted by their overall frequency and how exclusive they are to the topic).

To further evaluate topics' relevance, we report in Figure 5.2 the 15 topics sorted according to their prevalence, which represents the proportion of documents devoted to each topic. Market performance, Barclays and Deutsche Bank Libor manipulation and City group bailout represent the most relevant and covered topics showing a prevalence greater than 0.08.

The results obtained from weekly based STM with 15 topics appear quite consistent with those obtained from the monthly level analysis, easing the comparability of the final results. The additional topics highlighted by this analysis are: "Insider trading investigation" (Topic 17), "Wells Fargo-Wachovia acquisition" (Topic 18), "Bank management changes" (Topic 19) and "US banks stocks performance" (Topic weekly 20) in Table 5.4.

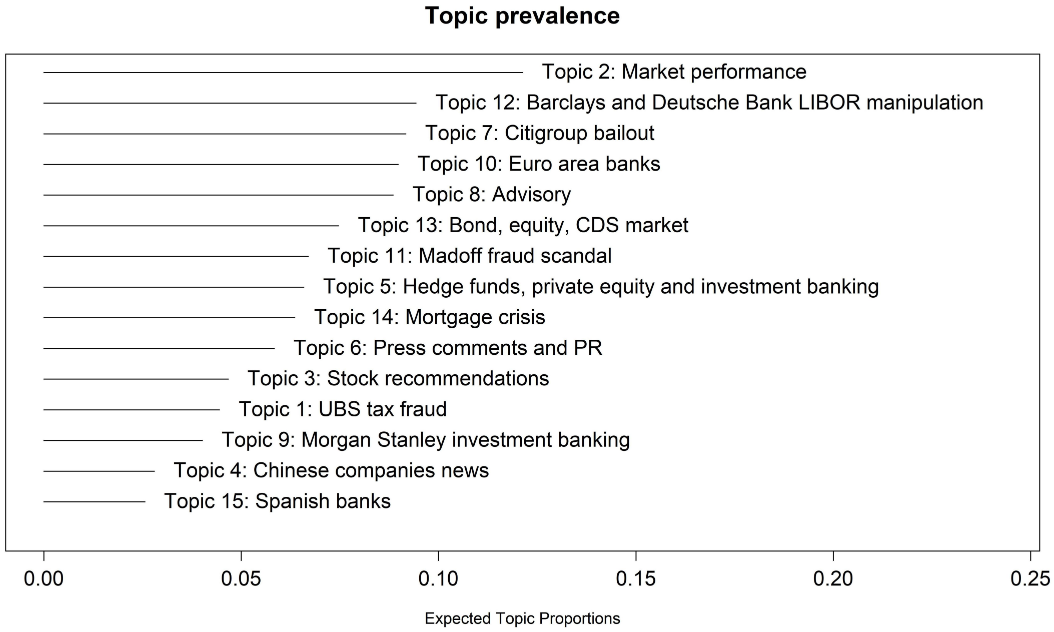


FIGURE 5.2: Topic prevalence the 15 topics analysis.

In Figures 5.3–5.11, we show how the two introduced covariates impact on the different topics in the two proposed scenarios (monthly and weekly based analysis).

Insofar, we can analyze either separately or in combination how the topics evolve through countries and time. In Table 5.6, we represent the topic proportions of the different topic in the different countries. Such analysis allows highlighting the specific country dependence of some topics such as the “UBS tax fraud scandal” upon Switzerland, the “Chinese companies news” upon China or the “Mortgage crisis” upon USA and UK. On the other hand, we can see topics more equally spread among the countries revealing a possible contagion/diffusion effect such as for “Madoff fraud scandal”, “Barclays and Deutsche Bank Libor manipulation” and “Citigroup bailout”.

Topic	**China**	**France**	**Germany**	**Spain**	**Switzerland**	**UK**	**USA**
UBS tax fraud scandal	0.01	0.03	0.04	0.01	0.13	0.03	0.03
Market performance	0.17	0.11	0.11	0.12	0.12	0.10	0.13
Stock recommend.	0.01	0.05	0.03	0.01	0.15	0.06	0.01
Chinese company news	0.43	0.01	0.01	0.01	0.01	0.03	0.01
H. Funds, Pr. Eq. and Inv. Bank.	0.01	0.01	0.01	0.01	0.01	0.01	0.12
Press comments and PR	0.02	0.02	0.03	0.02	0.03	0.02	0.09
Citigroup bailout	0.07	0.04	0.04	0.02	0.07	0.05	0.13
Advisory	0.03	0.07	0.20	0.03	0.10	0.14	0.06
Morgan St. Inv. Banking	0.00	0.01	0.01	0.00	0.01	0.01	0.07
Euro area banks	0.07	0.40	0.24	0.08	0.08	0.07	0.05
Madoff scandal	0.02	0.03	0.06	0.02	0.06	0.06	0.08
Barclays and DB LIBOR manip.	0.07	0.09	0.11	0.03	0.13	0.18	0.06
Bond, Equity and CDS markets	0.05	0.08	0.07	0.06	0.03	0.15	0.06
Mortgage crisis	0.04	0.03	0.04	0.02	0.03	0.07	0.08
Spanish banks	0.02	0.02	0.02	0.57	0.02	0.02	0.02

TABLE 5.6: Topic prevalence by country.

To consider jointly the temporal and spatial effect, we focus specifically on some interesting topics such as Topic 12 “Barclays and Deutsche Bank Libor manipulation”, Topic 10

“Euro area banks”, Topic 11 “Madoff fraud scandal” and Topic 14 “Mortgage crisis” that appear to be more diffused among the analyzed countries.

From Figures 5.3–5.6, we can gather insights regarding the topic proportions evolutions in monthly aggregated data over the different countries. For example in Figure 5.3, i.e., Topic 12 about “Libor manipulation”, a misalignment of the topic proportion peaks appearing for UK, Switzerland and Germany, suggests to further investigate through inferential tools. Similar considerations can be drawn for the other plots, for example, in Figure 5.5 for Topic 11 “Madoff scandal”, where the misalignment is evident for USA, Switzerland, Germany and France.

In Figures 5.7–5.11, we report the same plot analysis referred to the same topics now obtained through STM applied on weekly data. However, when performing the analysis at week level, we are left with four countries instead of five due to exclusion of some banks for the higher sparsity of the news data. Once again, we can observe different dynamics in the evolution of topics, particularly evident for “Madoff scandal” (Figure 5.8) and “UBS tax fraud” (Figure 5.11). At the same time, in Figure 5.7, regarding “Libor manipulation”, we can observe a different pattern compared to Figure 5.3: the whole topic depends more upon UK with particular turbulence during the weeks between March 2011 and March 2013.

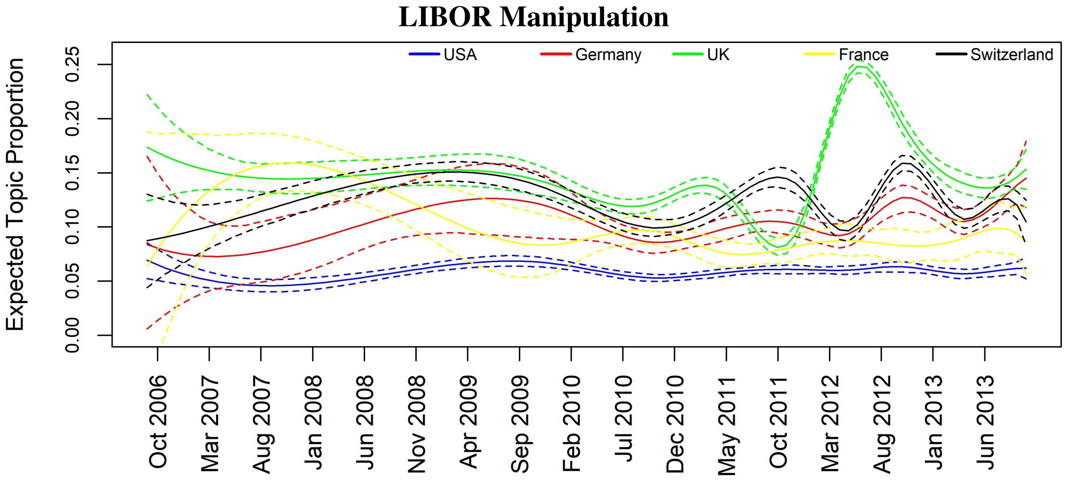


FIGURE 5.3: Topic prevalence evolution by country with respect to monthly based analysis.

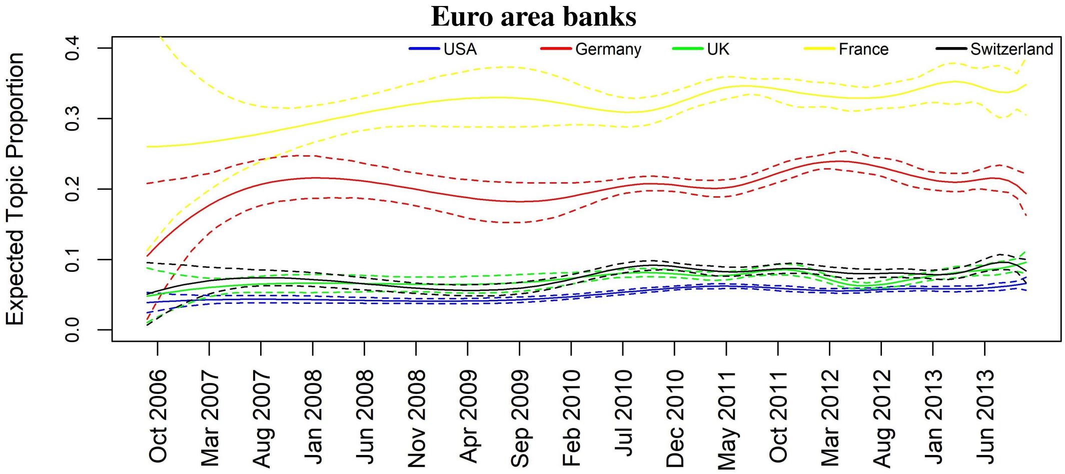


FIGURE 5.4: Topic prevalence evolution by country with respect to monthly based analysis.

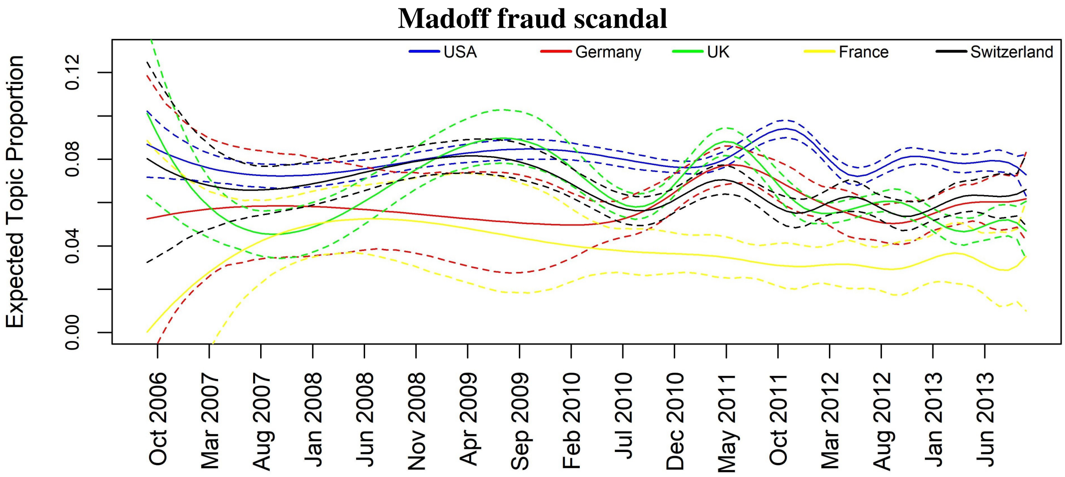


FIGURE 5.5: Topic prevalence evolution by country with respect to monthly based analysis.

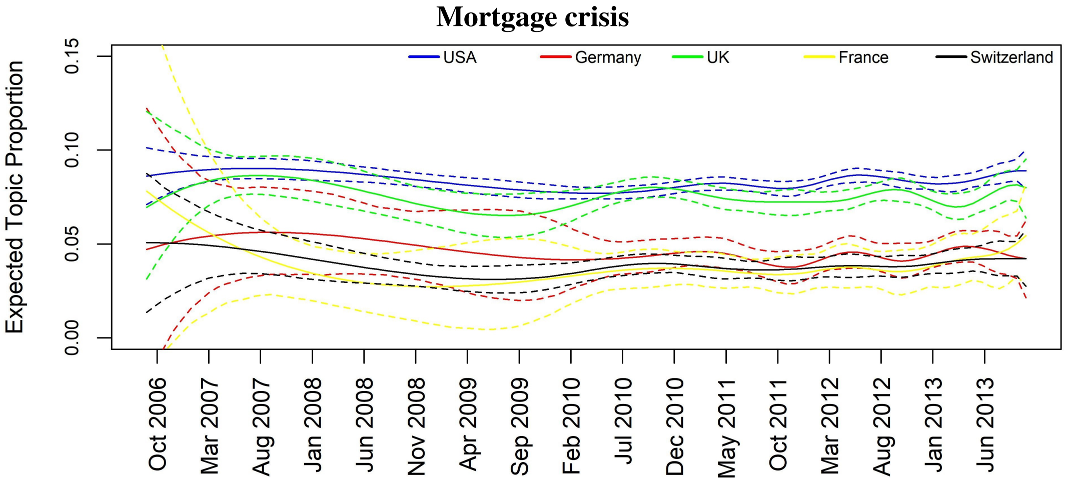


FIGURE 5.6: Topic prevalence evolution by country with respect to monthly based analysis.

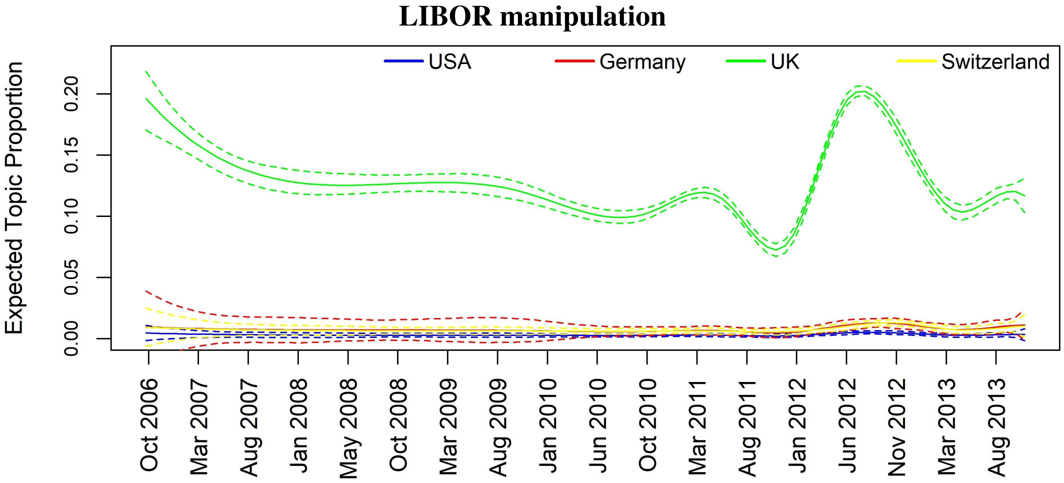


FIGURE 5.7: Topic prevalence evolution by country with respect to weekly based analysis.

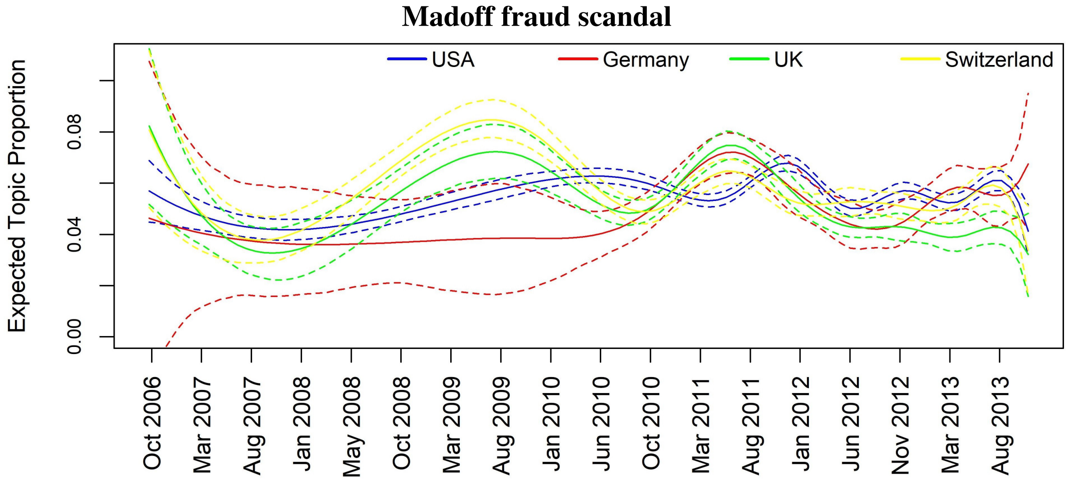


FIGURE 5.8: Topic prevalence evolution by country with respect to weekly based analysis.

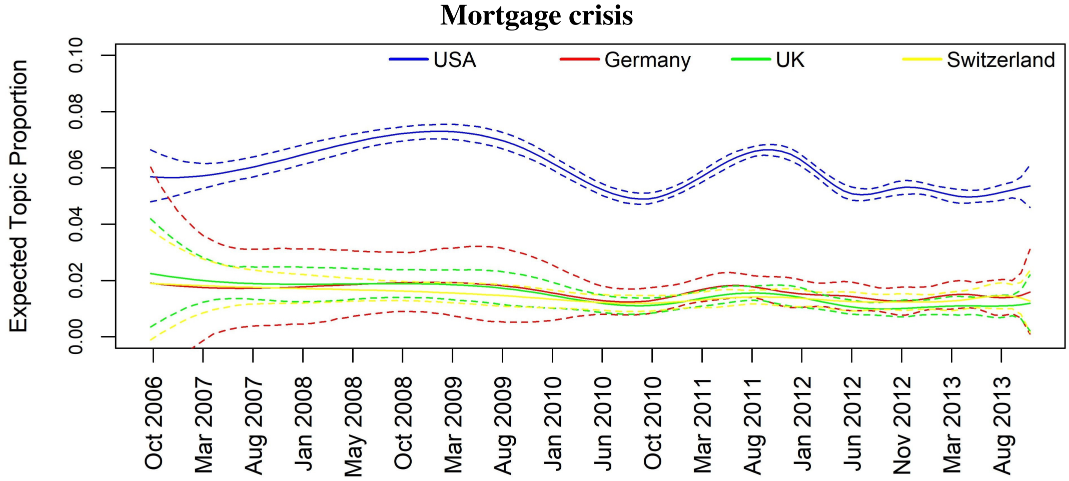


FIGURE 5.9: Topic prevalence evolution by country with respect to weekly based analysis.

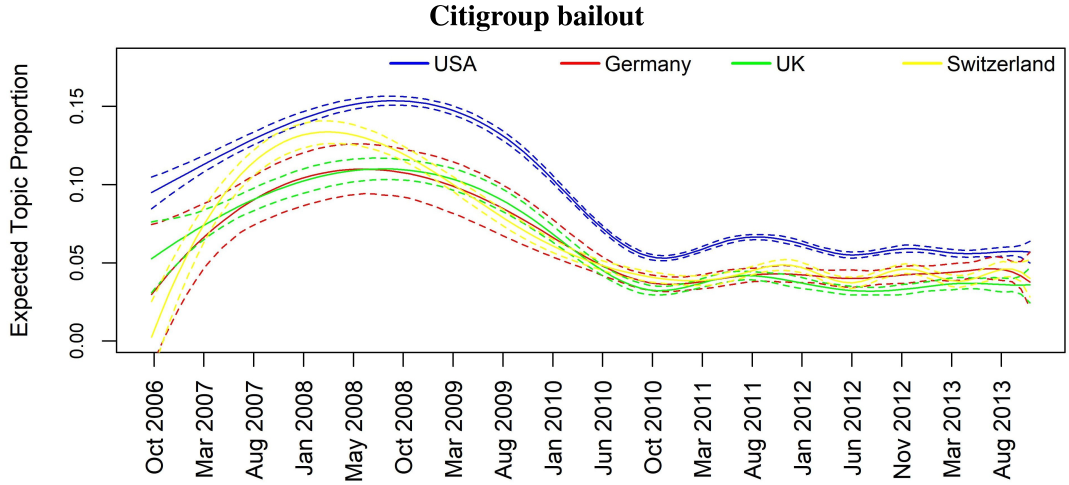


FIGURE 5.10: Topic prevalence evolution by country with respect to weekly based analysis.

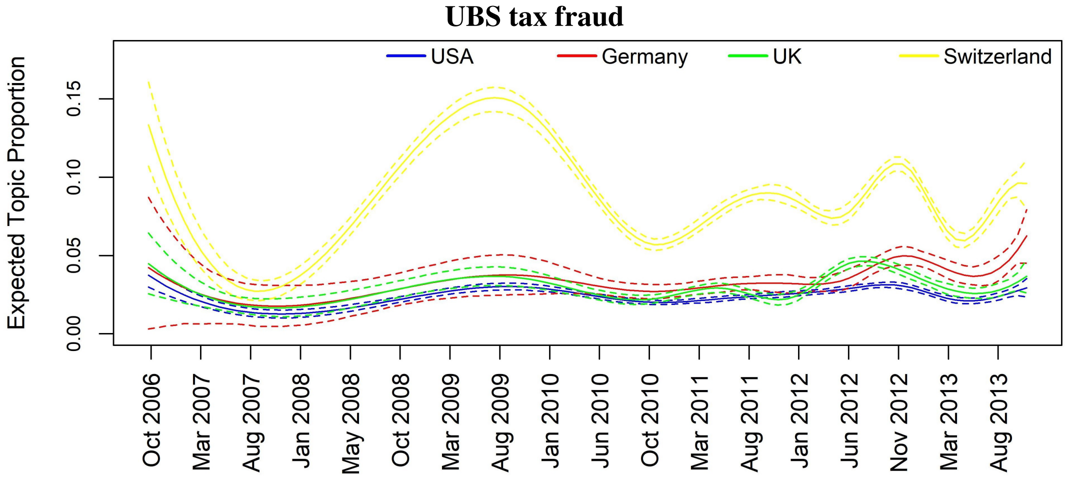


FIGURE 5.11: Topic prevalence evolution by country with respect to weekly based analysis.

Beyond the usefulness of a graphical inspection, we need an inferential tool, namely the Granger causality test, to possibly confirm our main research hypothesis: a given topic prevalent at time t in country c is also prevalent at time $t+1$ in country p according to a Granger causation influence.

Among the 15 discovered topics in the two analysis, we focus specifically on six arguments that we consider more important from a contagion point of view: "UBS fraud scandal (Topic 1)", "Citigroup bailout (Topic 7)", "Euro area banks (Topic 10)", "Madoff fraud scandal (Topic 11)", "Barclays and Deutsche Bank Libor Manipulation (Topic 12)" and "Mortgage crisis (Topic 14)". In Table 5.7, we include Granger causality results statistically significant at 5% for the topics listed above in the monthly based analysis, where 1L stands for one-month lag and similarly 2L for two-months lag. We can observe that there are several significant Granger causalities both at one and two-months lag. As one would expect, the Granger causation is present both within European countries and between USA and European countries, stressing the strict interconnection among countries from a financial perspective. We have excluded China and Japan from this analysis due to a limited number of available documents that can bias results (see Table 5.2).

UBS Tax Fraud	Significant Lag	Citigroup Bailout	Significant Lag
FR → USA	1L, 2L	FR → USA	1L, 2L
FR → UK	1L, 2L	CH → UK	1L, 2L
UK → DE	2L	FR → UK	1L
UK → FR	2L	USA → CH	1L, 2L
Euro Area Banks	**Significant Lag**	**Madoff Scandal**	**Significant Lag**
CH → USA	1L, 2L	UK → USA	1L, 2L
FR → USA	1L, 2L	CH → USA	1L, 2L
USA → UK	1L,2L	DE → UK	2L
CH → UK	1L,2L	DE → CH	2L
FR → UK	1L,2L	FRA → CH	2L
FR → CH	1L,2L	-	-
FR → DE	1L,2L	-	-
Libor Manipulation	**Significant Lag**	**Mortgage Crisis**	**Significant Lag**
CH → USA	2L	CH → USA	2L
CH → DE	1L	FR → UK	2L
-	-	USA → CH	1L, 2L
-	-	FR → CH	2L
-	-	USA → FR	1L, 2L
-	-	USA → DE	1L

TABLE 5.7: Results from Granger causality test for Topics 1, 7, 10, 11, 12 and 14 obtained from STM applied on monthly based data.

As examples, let us focus on results for Topic 11 (Madoff scandal) and Topic 14 (Mortgage crisis). Regarding the former, we can see that the influencing countries at one- and two-month lag are UK and CH whose banks had a high exposition towards the fraud, in particular HSBC, RBS and UBS. The importance of these two countries in the topic is justified from the fact that we are considering only banks' related news focusing primarily on the relation between banks and the fraud, and thus on the most exposed banks. In the Mortgage crisis, we can see how the information contagion is transmitted from USA to some European countries at one-month lag, namely FR, DE and CH (CH and FR also at two-months lag), and this is a plausible result as this specific financial crisis had origin in the United States. It is also interesting to give attention to Topic 10 regarding Euro area banks. All the interactions are significant at both one and two-month lag, and France seems to play a key role in spreading the topic among all the other European countries and USA.

Similarly, in Table 5.8, we include the Granger causality results statistically significant at 5% for the topics analysis based on weekly data, where 1L stands for one-week lag and similarly 2L for two-weeks lag.

UBS Tax Fraud	**Significant Lag**	**Citigroup Bailout**	**Significant Lag**
UK → USA	1L,2L	USA → UK	1L
USA → CH	1L,2L	USA → CH	1L, 2L
USA → DE	1L,2L	DE → UK	1L
UK → DE	2L	UK → USA	2L
-	-	DE → USA	2L
Mortgage Crisis	**Significant Lag**	**Madoff Scandal**	**Significant Lag**
UK → CH	1L, 2L	CH → USA	1L, 2L
CH → UK	1L, 2L	UK → CH	1L, 2L
USA → CH	2L	USA → DE	1L
-	-	UK → USA	2L
-	-	DE → USA	2L
-	-	UK → DE	2L
Libor Manipulation	**Significant Lag**		
CH → USA	1L	-	-
USA → UK	2L	-	-
CH → UK	2L	-	-

TABLE 5.8: Results from Granger causality test for Topics 1, 7, 14, 11 and 12 from STM applied on weekly based data.

Once again, we can see several significant results, rather similar to those in Table 5.7, although with some differences due to the different granularity of the data. If we look at "Madoff scandal", we can confirm the influence of UK and CH onto USA at both one and two-weeks lag and we further see a prominent role of UK in diffusing the topic. For the "Mortgage crisis", we have less evidence but we can confirm the contagion from USA to CH and the mutual causality between CH and UK. From Table 5.8, we can infer that even on more granular weekly data, although considering less countries because of news sparsity, we obtain several signs of causality for the most important and influential topics.

5.6 Conclusions

In this work, we have presented a fully data-driven methodology for the evaluation of news contagion through country and time dimension. We focused on SIFIs related news taken from two public dataset from Reuters News and Bloomberg News containing in total 553,666 documents spanning a period from October 2006 to November 2013. The aim of this study is to propose an approach for assessing the spread of news among countries along the considered time horizon. To this purpose, we have applied a model for topic modelling, called STM, able to fit the best topic distribution also on the basis of useful covariates that can be chosen by the analyst. The introduction of time and country specific variables has allowed us to add temporal and spatial dimensions to the analysis. This information have been exploited to investigate the dynamic of news spread among countries.

In particular, we have used the Granger causality test to demonstrate a contagion/causation dynamic in the diffusion of the news employing topic proportion timeseries extrapolated from the STM approach. Such analysis has been conducted considering two different data granularities: news aggregated on a monthly basis or on a weekly basis. According to the two different time references, it is necessary to reduce the list of considered banks (and associated countries) to have enough data coverage for fitting reliably the STM model. Whilst we have analyzed weekly data, for some country/bank combinations, we are left with a insufficient data coverage, possibly producing a bias in the results that should be taken into account while comparing them with the monthly based analysis.

In both cases results are promising, we have found several significant causal relations in the diffusion of the news, stimulating further development in a future work. In particular, we shall investigate a correlation structure in the news diffusion taking into account country or bank level with correlation network models. Moreover, the analysis should be conducted with a more populated dataset, ideally the full Reuters and Bloomberg corpus from to 2007 to 2015 to increase the list of considered banks, and thus producing even more detailed and insightful results.

Chapter 6

Twitter Sentiment and Banks' Financial Ratios: Is There Any Causal Link?

6.1 Summary

In this chapter we study the relationships between Twitter sentiment and various financial indicators (e.g. stock returns or trading volume) of some of the major Italian banks. Moreover, we test the current technology for analyzing and evaluating the sentiment of short web-text messages written in Italian, such as those published on the Twitter micro-blogging platform.

In fact, gauging the sentiment among financial investors is of paramount importance for both market participants and regulation authorities. Behavioural finance posits that stock market investors define their purchasing strategies considering arbitrage bounds and collective sentiments. Regulation and market authorities can address critical situations by collecting and analyzing the sentiment mood inferred from investors' actions on social media.

Here our goal is to establish a statistical framework to measure the causal links between sentiment extracted from Twitter and financial market variables even in presence of no stationarity and cointegration in the data.

A quantitative evaluation of the impact of sentiment on financial indicators is relevant to increase the timely awareness of regulators with respect to potentially critical microeconomic conditions.

6.2 Introduction

The sheer amount of detailed on-line information streaming from social networking and micro-blogging platforms such as Twitter, is increasingly attracting the attention of researchers and practitioners from many different fields. The linguistic analysis of social media contents has become a hot topic even for applied research in different languages. Detection of sentiments and opinions in social media is now a critical tool for monitoring social media platforms.

As a matter of fact, social media contents constitute a relevant asset for private firms and public institutions to tap into the customers' needs and preferences in real time. Insofar, pulling out sentiments from Twitter has been already employed for several purposes: to monitor political sentiment [Tumasjann et al. 2010], to extract critical information during times of mass emergency [Tumasjann et al. 2012], to check the effects of sentiments on stock price and volume in the Dow Jones Index [Ranco et al. 2015] or on market share in the Italian financial market [Cerchiello and Giudici 2015].

Indeed, in April 2013 the Securities and Exchange Commission (SEC) issued a report that allowed companies to use social media outlets like Facebook and Twitter to announce key information in compliance with Regulation Fair Disclosure as long as investors had been notified about which social media platform would have been used to disseminate such information.

Today the use of micro-blogging platform like Twitter has gained a sound position for both the US and the UK markets [Mao et al. 2015]. Even in Italy the use of Twitter among financial practitioners has grown steadily in the last five years.[1]

Although Twitter has already gained a solid reputation as information source in the United States, the situation is still unclear for Italy and other European countries in general. In fact the level of empirical correlation between financial time series and Twitter-derived sentiment has not been deeply investigated yet. We believe that a similar task for Italian language and the development of a standard sentiment corpus will foster a better understanding of how sentiment is conveyed in tweets. Training and testing automatic systems obviously require the availability of several resources that may consist in large datasets of annotated posts or even in lexical databases where affective words are associated with polarity values.

In this chapter, we focus on tweets written in Italian and obtained from Twitter related to some of the most important Italian banks: Intesa San Paolo (ISP), Monte dei Paschi di Siena (BMPS) and Unicredit (UCG). The share of these banks in total market capitalization amounts to around 70%, while in terms of total assets accounts for around 80% of the Italian listed banks. For robustness purposes we also analyze Deutsche Bank (DBK).

The main task concerns the evaluation of sentiment polarity classification at the tweet message-level. Successively, these message-level measures are aggregated on a daily basis. Sentiment expressed in tweets is usually classified as i) positive, ii) negative, or iii) neutral. However a message can contain parts expressing both positive and negative sentiments (therefore a mixed sentiment), a feature that should be somehow tackled.

Considering that the availability of pre-labelled text for Italian is currently very limited [Basile and Nissim 2013, Bosco et al. 2013], we try to extract the sentiment through a dictionary-based approach that maps preassigned lists of positive and negative words onto the collected tweets. The final score is then given as an appropriate function of positive and negative counts. Specifically we have investigated the library TextWiller written in R[2]. Such library is developed for unsupervised sentiment analysis and presents a list of specific words in Italian with both positive and negative polarities. Thus, the sentiment classifier is based on those words and accounts for the relative quotas of positive and negative words in each Twitter message.

This work contributes to the recent burgeoning literature on social media and financial markets. There is already a number of papers that use Twitter data to generate new sentiment measures and correlate them with financial figures [Sprenger and Welpe 2010, Bollen et al. 2011, Mittal and Goel 2012, Rao and Srivastava 2012, Nann et al. 2013, Oliveira et al. 2013]. There is another strand of literature that uses social media as an alternative way to release information, thus reducing information asymmetry and improving stock liquidity, attracting more investors. Other papers such as [Chawla et al. 2016, Giannini et al. 2013] use Twitter data dynamically to see how information diffusion affects trading and how track changes in investor disagreement. It also connects to a large literature on retail trader' attention [Barber and Odean 2008].

[1] According to http://www.internetlivestats.com/internet-users/italy/ internet users in Italy in 2016 were around 39 millions, i.e. around 66% of the total population. According to https://www.statista.com/statistics/260721/number-of-social-network-users-in-italy/ the number of social network users in Italy in 2014 was 21.6 millions. According to http://www.digitalnewsreport.org/survey/2015/italy-2015/ the popularity of Twitter in Italy is still marginal as that of Google+ when compared to Facebook, which is the dominant social network platform around the globe. In fact, for example only 10% of the Italians use Twitter weekly for searching news and we can find similar statistics for the UK or the US.

[2] see https://www.r-project.org/

The rest of chapter is organized as follows. After the introduction, in section 6.3 we present the basic theoretical concepts behind the text-mining and sentiment analysis techniques we have adopted and about Granger causality text. Section 6.4 describes how the data has been collected, cleaned and prepared for the analysis. Section 6.5 reports our main findings in terms of significance of the several causality tests applied to data at hand. Finally section 6.6 provides some discussion and concluding remarks.

6.3 Methodology

The approach proposed in this chapter is an appropriate combination of a number of methods necessary for cleaning up the corpus, extracting reliable sentiment from it and for finally employing such sentiment to prove its possible causal effect onto standard financial figures. We suggest such strategy as a kind of protocol for tagging Italian text, specifically Italian tweets, with regards to the expressed sentiment and then using such information for assessing a causal link to banks' financial ratios.

6.3.1 Sentiment Analysis

Quantitative analysis of human languages allows to discover common features of spoken or written text. In particular the analysis of short text messages like those appearing on microblogging platform presents a number of challenges. Some of these are, the tweets length limit initially forced by the social media platform to 140 characters and later relaxed till 280, the informal conversation (e.g. slang words, repeated letters, emoticons) and the level of implied knowledge necessary to understand the tweets. Moreover, it is important to consider the high level of noise contained in the tweets, witnessed by the fact that only a small fraction of them with respect to the total number available is employed in our sentiment analysis.

This selection has been carried out in two steps. The first step consists of an exploratory analysis for understanding the dataset while the second step selects the tweets on a keywords basis. For the exploratory part we considered an unsupervised clustering procedure based on a combination of text vectorization, Latent Semantic Analysis (LSA) and *k-means* clustering. We have applied a Bag of Word (BoW) approach, where a text is represented as an unordered collection of words, considering only their counts in each tweet. The word and document vectorization has been carried out by collecting all the word frequencies in a Term Document Matrix (TDM). Afterwards such matrix has been weighted by employing the popular TF-IDF (Term Frequency Inverse Document Frequency) algorithm. LSA [Deerwester et al. 1990] is a methodology which applies Singular Value Decomposition (SVD) to the TDM. The following step consists in picking a threshold below which all the singular values are replaced by zero. In this way we can reduce the dimensionality of the vector space where documents are embedded. This space is equipped with an Euclidean measure which allows us to evaluate the distance among documents. Finally, to group together similar documents, we applied k-means clustering on this lower-dimensional space. The subsequent document inspection by sampling (casual or according to distance from the centroid) can be very insightful in understanding documents topics and in identifying which clusters contain relevant and irrelevant documents.

In addition to these methodologies, Latent Dirichlet Allocation LDA, [Blei et al. 2003], has been used to investigate the main topics of discussion for each bank in different periods and see how the different topics prevalence changes over time. For this task, LDA is very useful since it allows to inspect the topics both within each specific cluster and within the entire collection of tweets. The characteristics of topics emerging from LDA, in terms of differentiation among each others, provides also a feedback on the number of clusters employed in the k-means algorithm. In Figures 6.1 and 6.2 it is possible to see the time-evolution of

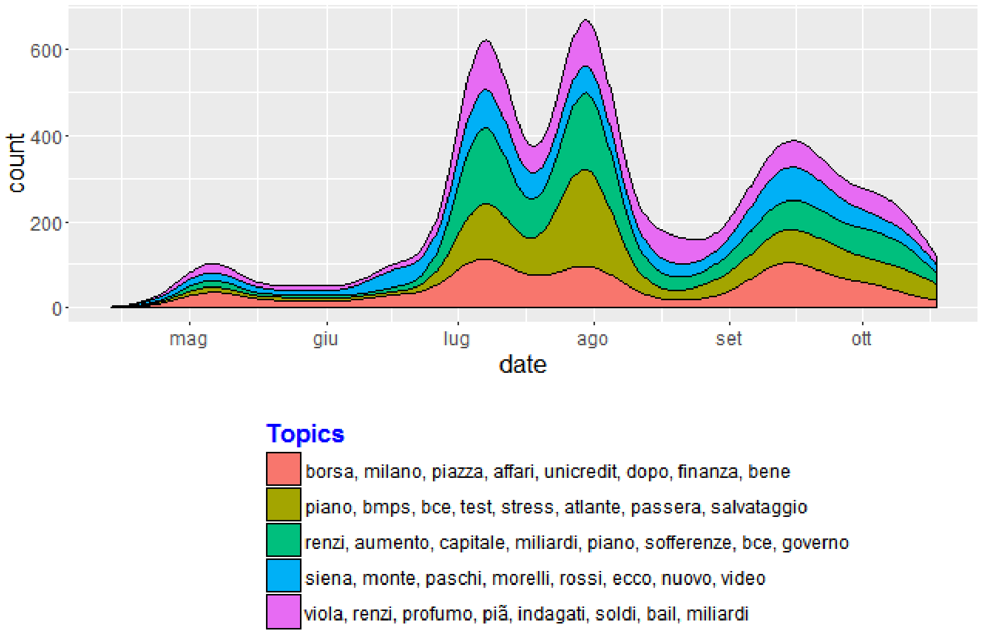


FIGURE 6.1: BMPS absolute topic prevalence

topic prevalence for MPS measured respectively in terms of number of tweets and percentage of tweets. From the LDA is possible to identify 5 topics of discussion covering the market sentiment about the bank, the private rescue plan of MPS trough the Atlante fund, the government backed rescue plan, the fatal accident occurred to a member of the management and a more general topic about bank stress.

In the second step, in order to select only the relevant tweets we have applied a keyword based approach, restricting the keywords on the results of the previous analysis, after excluding the retweets. Due to the high number of tweets scraped (for both keyword selection and retweets exclusion), a time consuming manual inspection can be applied only on few documents and automated methods like LSA and LDA are necessary both in the collection (scraping) and the selection of relevant tweets. Through such analysis, we are able to identify which tweets are relevant or not to our problem and we can then proceed with the analysis.

The most critical part of the analysis relies on the sentiment classification. In general, two different approaches can be used:

- Score dictionary based: the sentiment score is based on the number of matches between pre-defined list of positive and negative words and terms contained in each text source (a tweet, a sentence, a whole paragraph);
- Score classifier based: a proper statistical classifier is trained on a large enough dataset of pre-labeled examples and then used to predict the sentiment class of a new example.

However, the second option is rarely feasible for less widespread languages like Italian, because in order to fit a good classifier, a large amount of pre-classified examples is needed. This represents a particularly complicated task when dealing with short and extremely non conventional text like tweets. Insofar, we decided to focus on a dictionary based approach, building appropriate lists of positive and negative words relevant for financial topics in Italian language. The basic vocabulary is based on that available within R package TextWiller. The positive list comprises 980 stems (root of a word) and the negative one 2277. However such lists are non content specific, thus we have enriched them by adding some relevant words as follows:

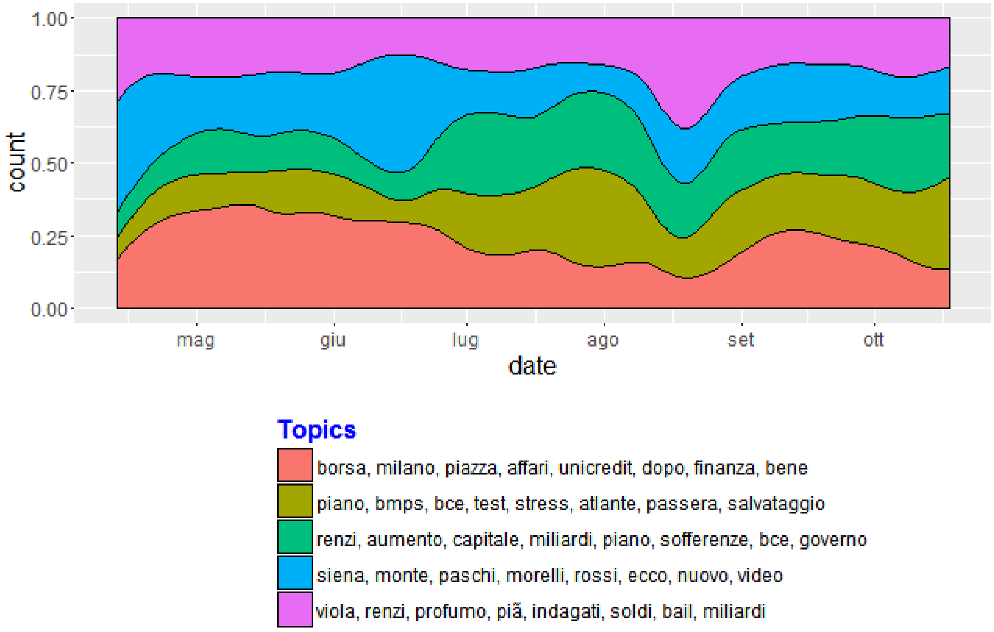


FIGURE 6.2: BMPS relative topic prevalence

- New negative words = tonfo, calo, sofferente, tracollo, rimettere, rosso, scendere, minimo, massone, caduta, mafia, ciclone, picco, dilapidare
- New positive words= salire, profitti, rialzo

According to the number of matches between the terms contained along the tweets and the above defined list of positive/negative words, the Twitter Sentiment Index (hereafter TSI) is calculated as follows:

$$TSI = \frac{\#PW - \#NW}{\#PW + \#NW} \tag{6.1}$$

where #PW represents the number of positive words matched and similarly #NW the number of negative words matched. Therefore, the index TSI is calculated as the ratio between the excesses of positive words with regards to negative words and the total number of positive and negative words. The denominator is useful not only to obtain a relative number but also to control for possible biases due to the different length of the tweets (remember that since November 2017, Twitter increased the maximum number of characters from 140 to 280).

Thanks to this approach we were able to classify the polarity of ca. 325,000 tweets (without considering re-tweets). As a result, we get a TSI ranging from -1 to +1 for each and every available tweet. Then, an appropriate merging strategy has been used to get a final score associated to each bank-day combination. First, the daily twitter volume $TweetVol_i$ has been computed by counting the number of tweets in each day, obtaining a timeseries for each bank with the daily tweet volume.

Then a timeseries with the average TSI of the day ($Sent_{ts_i}$) has been also computed for each bank. The result is a timeseries recording the day-by-day evolution of the daily average TSI. The measure is computed as follows, dividing the sum of the all the TSI of one day by the Tweet volume of that day ($TweetVol_i$):

$$Daily_{TSI_i} = \frac{\sum_{j=i} TSI_j}{TweetVol_i} \tag{6.2}$$

where i is the considered day and the equation is reported for a single bank.

In addition, we propose a simple Twitter sentiment measure that takes into account also the daily tweet volume combining the two information together. We will refer to it in the rest of the chapter as "Twitter sentiment weighted" ($Sent_{We}$). The new quantity, is obtained simply by taking the sum of all the TSI in one day and dividing it by the average daily tweet volume over the entire observation period for each bank. In practice this allows to weight more days with high spikes of tweets polarized in a certain direction as it happens often in proximity of important events and news disclosures. Dividing by the average $TweetVol$ over the observation period is done for normalization purpose.

$$Sent_{We_i} = \frac{\sum_{j=i} TSI_j}{\overline{TweetVol}} \tag{6.3}$$

where i is the considered day, $\overline{TweetVol}$ is the average daily tweet volume over the observation period and the equation is reported for a single bank.

6.3.2 Granger Causality

In order to evaluate the size and the direction of the possible causal relationships between our financial variables (such as stock return, traded volume, etc.) on one side and the sentiment indicator on the other side, we firstly test the stationarity of each time series and then we apply the Granger causality test [Granger 1969]. The stationarity of each time series has to be tested by applying the following Augmented Dickey-Fuller regressions:

$$\Delta(y_t) = (\rho - 1) \cdot y_{t-1} + \sum_{i=1}^{p} \beta_i \cdot \Delta(y_{t-i}) + \varepsilon_{t-1}^{1} \tag{6.4}$$

$$\Delta(y_t) = Cnst + (\rho - 1) \cdot y_{t-1} + \sum_{i=1}^{p} \beta_i \cdot \Delta(y_{t-i}) + \varepsilon_{t-1}^{2} \tag{6.5}$$

$$\Delta(y_t) = Cnst + \gamma \cdot trnd_t + (\rho - 1) * y_{t-1} + \sum_{i=1}^{p} \beta_i \cdot \Delta(y_{t-i}) + \varepsilon_{t-1}^{3} \tag{6.6}$$

where: y_t is the series being tested for stationarity; $\Delta(y_t) = (y_t - y_{t-1})$ is the time first difference. The three forms represent, respectively: a random walk (Equation 6.4), a random walk with drift (Equation 6.5) and finally a random walk with drift and a linear trend (Equation 6.6).

In all of the three previous equations the null hypothesis (H_0) subject to test is the nullity of the coefficient $(\rho - 1)$ and the test statistic is given by the t-statistics on such coefficient. When this coefficient is close to zero we cannot reject the null hypothesis of independence[3].

Lack of stationarity in the set of variables being tested for Granger's causality would cause misspecification in the tests (see [Granger 1988] for further details). We performed an Augmented Dickey-Fuller (ADF) tests for twitter derived timeseries and for the financial variables timeseries (e.g. stock return, CDS and Bond spreads, etc.). For example, from the results of the Augmented Dickey Fuller tests we reject the hypothesis of the presence of a unit root in time series of the Twitter Sentiment Index and the stock return for the bank BMPS. Given the results of the ADF tests and the stationarity of the time series involved, there is no need to run cointegration tests. The same ADF tests have been carried out on the other

[3] Independence is lack of Granger causality in either direction.

timeseries for all the banks under investigation: Intesa San Paolo (ISP), Unicredit (UCG), Monte dei Paschi di Siena (BMPS) and Deutsche Bank (DBK).

The Econometrica Granger's paper 'Investigating Causal relationship' [Granger 1969] has sparkled a huge literature on the micro as well as macro econometric models. The Granger causality principle is straightforward: if lagged values of X_t contribute to foresee current values of Y_t in a forecast achieved with lagged values of both X_t and Y_t then we say *X Granger causes* Y_t. As was first shown by Sims [Sims 1972], the Granger causality corresponds to the concept of exogeneity and it is therefore necessary to have a unidirectional causality in order to guarantee consistent estimation of distributed lag models.

In our empirical experiment we have considered the following equation:

$$y_{t+i} = \beta_0 y_t + \beta_1^y x_t + e_{t+i} \tag{6.7}$$

$$x_{t+i} = \beta_0 x_t + \beta_1^x y_t + e_{t+i} \tag{6.8}$$

where we want to test whether x_t Granger causes y_t and vice versa. Our null hypothesis is therefore: $H_0 : \beta_1^y = \beta_1^x = 0$. Taking into account that we are dealing with daily time series, in our tests we have considered up to five lags to take into account the effect of a business week.

Table 6.2 shows the final results of the Granger causality tests for the four banks included in our sample (BMPS, ISP, UCG, and DBK) when both a constant and trend are included. In Table 6.2 the Daily Twitter sentiment score ($Daily_{TSI}$) is computed as the simple average of the sentiment computed as explained in the previous section. For each bank we compute the Granger causality test for six financial variable: i) stock returns (r_t), ii) volume of stocks traded (VV_t), iii) Volatility (Vol_t) computed as the daily range, iv) senior CDS spreads (CDS_t^{Sen}), v) subordinated CDS spreads (CDS_t^{Sub}), and vi) an average subordinated bond spreads ($Bond_t^{Sub}$).

6.4 Data

Using the public API (Application Programming Interface) provided by Twitter, we collected all tweets and retweets in Italian from all the active Twitter accounts which contain either the name, the hashtag or the ticker of a restricted set of Italian banks. In particular, we collected all the tweets and retweets for i) Monte dei Paschi di Siena (ticker: BMPS), ii) Unicredit (ticker: UCG) and iii) Intesa San Paolo (ticker: ISP). We also collected the tweets and retweets for the German bank Deutsche Bank (ticker: DBK) to check the robustness of our results with a non-Italian bank which was heavily cited in the news in the investigated period. We have been tracking such Twitter activity about Italian Banks and DBK for more than 28 months from August 2015 through January 2018. The relevant descriptive statistics about these tweets are summarized in Table 6.3 where we report the bank name, the ticker, the number of total tweets downloaded, the number of retweets, the number of tweets in Italian, the number of tweets used in our analysis, the average daily number of used tweets. We can notice that by far the highest number of total tweets was downloaded for DBK but this just because both English and Italian tweets were collected. In fact, if we restrict the study only to the Italian ones, we end up with around 79,000 tweets for DBK and it emerges clearly that the most covered bank is BMPS (around 606,000). However, this extremely high number of tweets is not entirely related to the recent problems faced by the oldest bank in the world that ended up with a precautionary recapitalization in December 2016. In fact, tweets related to BMPS were searched including the keywords "*MPS*" or "*Banca Monte dei Paschi*". However, we noticed that the former keyword resulted in a lot of English tweets related to the UK 'Members of Parliament' whose acronym is indeed 'MPs'. Actually, the Twitter API does not differentiate between lower and upper case and therefore 70% of tweets for BMPS ended up being tweets related to the Brexit discussion by the MPs of the British Parliament. The methodology adopted in the tweet pre-screening based on BOW, LSA and clustering has been very helpful in identifying these phenomena. Furthermore, still from Table 6.3 we can notice that 55% of the Italian tweets are just retweets by other Twitter users. In the end, for 28 months we downloaded around 780,000 tweets in Italian and after a typical cleaning we were left with around 260,000 tweets for BMPS, with an average of 341 tweets per day. For the other banks in our sample, we did not face similar problems with the keywords. For example, for UCG we downloaded around 27,000 tweets, 23,000 of those in Italian with around 4,000 retweets. The final number of tweets used for UCG was about 18,000, i.e. around 25 per day. For ISP we downloaded just 14,000 tweets that were mostly in Italian (12,000) with a daily average of 16.

All financial variables were downloaded from Bloomberg. Figures 6.3 to 6.6 depict the time series of the number of daily tweets in Italian regarding that bank, the simple daily average sentiment ($Sent_{ts}$) on those tweets and the Twitter sentiment weighted ($Sent_{we}$) on the same tweets. The latter could also be interpreted (in alternative to Equation 6.3) as the simple daily average sentiment ($Sent_{ts}$) weighted by the ratio between the number of tweets about that bank on each day and the daily average number of tweets over the observation period. As we can notice, for BMPS the simple average of Twitter Sentiment varies between -1 (negative) and 1 (positive), while the weighted average Twitter sentiment has some negative spikes that go beyond -1 in those days in which there is a lot of activity on the Twitter social platform. It is to note that for the two banks that faced more discussions in the news and reputational issues among the sample (i.e. BMPS and DBK) there have been higher negative spikes in the Twitter sentiment weighted ($Sent_{we}$). These spikes are due to those days where a particular negative news was heavily tweeted and thus the number of negative polarized daily tweets was higher than the daily over the observation period. For the other banks this does not seem to be the case and we can notice more positive spikes. In the end the effect of the weighted average sentiment is to give more importance to those days when a lot (compared to the average daily number of tweets over the observation period) of positive or negative tweets have been posted.

Figures 6.4, 6.6, 6.8 and 6.10 show the time series of the six financial variables we are analyzing for the four banks of our sample. Each bank in our sample has its own features due to the particular events and news that were spread during our sample period. The corresponding time series of each financial variable have also daily frequency and behave accordingly to what previously described for the Twitter timeseries.

6.5 Results

We include the main findings from our analysis in Table 6.2 and 6.4. On the left panel Table of 6.2 we test causality from financial variables to $Sent_{ts}$ on that bank, while on the right panel we test causality from $Sent_{ts}$ to the financial variable in each row. The Twitter sentiment $Sent_{ts}$ is computed as the simple average of sentiment on that bank. The stars *, **, and *** indicate rejection of the null hypothesis of no Granger causality at 5%, 1%, and 0.1%, respectively. All tests in both directions are computed from 1 to 5 lags to take into account the effects within an entire business week.

As we can notice, for BMPS, ISP and DBK we accept the null hypothesis of no Granger causality from Twitter Sentiment to all the financial variables. This means that the daily average Twitter sentiment alone does not Granger cause the financial variables for these banks from lags 1 to 5. Only for UCG we reject the null hypothesis of no Granger causality from the daily average Twitter Sentiment to the Subordinated CDS for lags 4 and 5. In the left panel we can also see that there is some Granger causality from financial variable to Twitter sentiment which seems to point in the direction of suggesting that financial variables drive the mood of the discussion on Twitter. If there are some news or events which are particularly important to a bank, these should spread out on the web and social platform causing some chattering among the web users and inspiring some form of agreement or disagreement which would be reflected on Twitter. We can find this for some variables (especially volume, CDS spreads and subordinated bond spread) and across all banks in our sample. In Table 6.4 the Twitter sentiment weighted $Sent_{we}$ is computed for each bank as the sum of tweets sentiment in each day divided by the average tweets volume over the observation period (as shown in Equation 6.3). Here we can notice that we do find a much higher number of variables which are Granger-caused by the Twitter sentiment weighted, suggesting that the tweets volume plays a relevant role. For BMPS, the stock return, volume and CDS^{sub} at all lags, and CDS^{sen} for all lags except for the second and the third are Granger caused by the Twitter sentiment weighted. For ISP the subordinated Bonds at all lags show Granger causality from the Twitter sentiment. The Twitter sentiment weighted seem also to Granger cause volatility and CDS spreads for UCG. For DBK, the returns at all lags and traded volume at lags 3 - 4 - 5, and the subordinated bond spread at lag 5. We also find several Granger causalities in the opposite direction (from financial variables to the Twitter sentiment weighted) for all the four banks. In particular for: the subordinated bond spread of BMPS; returns, volume, volatility and subordinated CDS spreads of ISP; the CDS spreads and subordinated bond spread of UCG; the CDS spreads of DBK.

All the results presented so far indicate a clear rejection for some financial variables and some banks of the null hypotheses in both directions. On one hand, we reject the null that the Twitter sentiment weighted does not Granger cause the financial variables at some lags. On the other hand, we reject the null hypothesis that the financial variable does not Granger cause the Twitter sentiment (weighted and not) at some lags. Therefore, we find some statistical evidence that we have a closed-loop feedback relationship between the two sets of variables.

6.6 Conclusions

In this chapter we have analyzed how a sentiment measure computed using data from social media platforms such as Twitter affects some financial variables (stock returns, volume, volatility, senior and subordinated CDS spreads, subordinated bond spreads) related to Italian banks.

In particular, starting from tweets written in Italian, we have shown how to extract a proxy for the sentiment conveyed by the short text messages present in the Twitter micro-blogging platform.[4]

We have also investigated how the volume of tweets can affect the same financial variables. We have first tested for the stationarity of the variables and then we proceeded with the Granger causality test to disentangle whether financial variables affects Twitter sentiment or vice versa. From our Granger causality results it seems that sentiment does Granger cause some financial variables for some Italian banks, while it is more often the case that financial variables do Granger cause sentiment on Twitter.

Our results show that both Twitter sentiment and Twitter volume do significantly affect some financial variables (such as stock returns, volatilities, volume or CDS spreads) for some of the banks in our sample. In particular, they affect several financial variables of BMPS and

[4] By taking advantage of some packages available on the CRAN (Comprehensive R Archive Network, https://cran-r-project.org repository), we have written some *R* procedures implementing different algorithms for text mining and sentiment analysis.

DBK which have recently experienced many episodes of high volatility and negative news and UCG.

We improve upon the previous literature by checking the correlation of more financial variables with respect to Twitter sentiment and volume. So far in the literature most of the papers have focused on stock returns posing less attention to different markets such as the credit or bond market of the investigated security.

We believe that our results are important for testing economic and finance theories and for policy purposes.

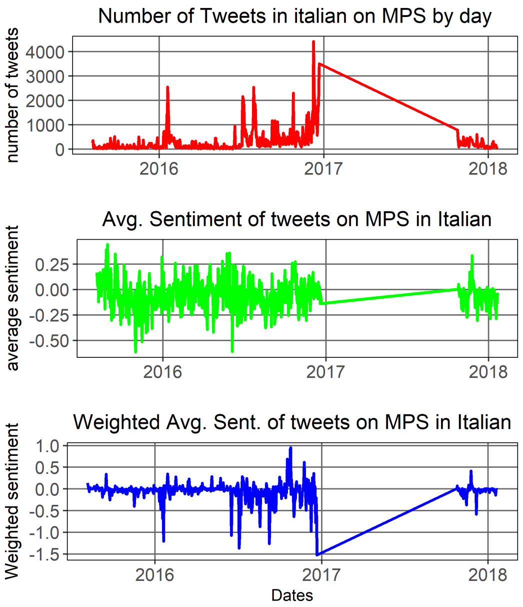


FIGURE 6.3: BMPS Twitter data

Notes: The figure depicts the number of tweets in Italian for bank BMPS on the top panel. The medium panel depicts the average sentiment of tweets on bank BMPS computed without any weightings on positive and negative tweets. The bottom panel depicts the weighted average sentiment of tweets on bank BMPS where the weight is given by the daily number of tweets divided by the average number of tweets in the whole sample.

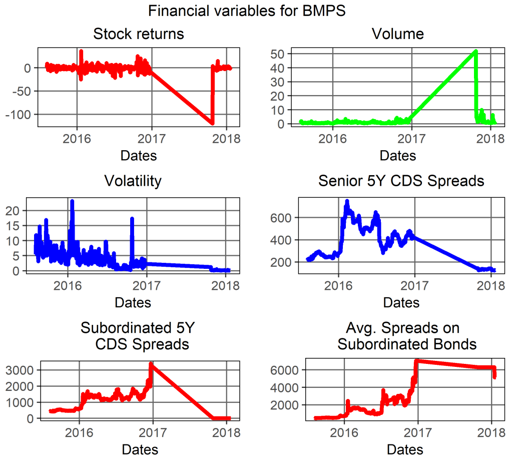


FIGURE 6.4: BMPS financial data

Notes: The figure depicts for bank BMPS from upper left panel to lower right panel: 1) stock returns, 2) volume of stocks traded, 3) volatility of stock returns, computed as the daily range (i.e. the difference between the highest and lowest price within each day), 4) the spread on the 5-year senior CDS, 5) the spread on the 5-year subordinated CDS, 6) the average spread on a selection of subordinated bonds issued by the bank.

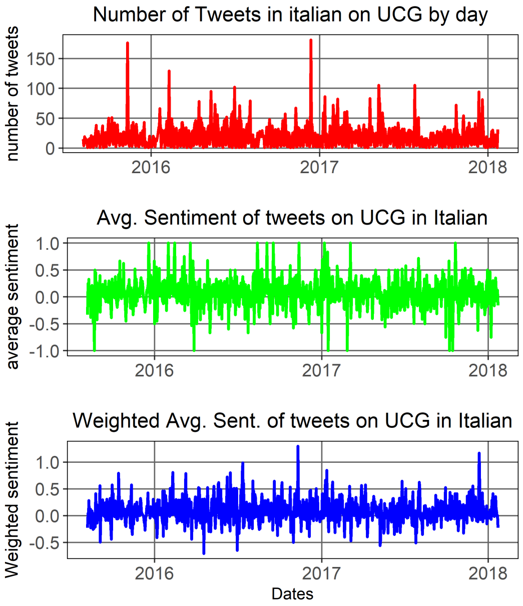


FIGURE 6.5: UCG Twitter data

Notes: The figure depicts the number of tweets in Italian for bank UCG on the top panel. The medium panel depicts the average sentiment of tweets on bank UCG computed without any weightings on positive and negative tweets. The bottom panel depicts the weighted average sentiment of tweets on bank UCG where the weight is given by the daily number of tweets divided by the average number of tweets in the whole sample.

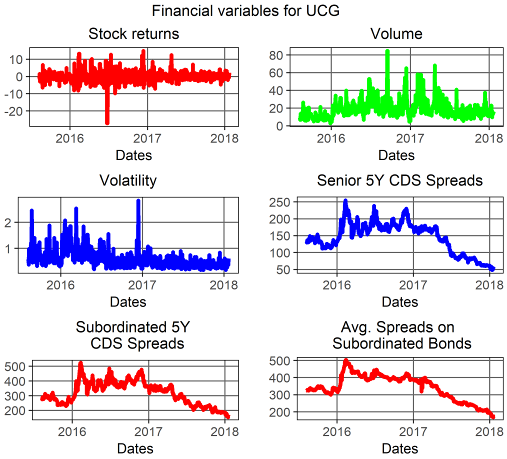


FIGURE 6.6: UCG financial data

Notes: The figure depicts for bank UCG from upper left panel to lower right panel: 1) stock returns, 2) volume of stocks traded, 3) volatility of stock returns, computed as the daily range (i.e. the difference between the highest and lowest price within each day), 4) the spread on the 5-year senior CDS, 5) the spread on the 5-year subordinated CDS, 6) the average spread on a selection of subordinated bonds issued by the bank.

Bank	**Ticker**	**Keywords**
Monte dei Paschi di Siena	BMPS	‘MPS’, ‘Banca Monte dei Paschi di Siena
Unicredit	UCG	‘Unicredit’
Intesa San Paolo	ISP	‘Intesa San Paolo’, ‘Banca Intesa’
Deutsche Bank	DBK	‘Deutsche Bank’

TABLE 6.1: Keywords used to download the tweets for each bank from Twitter

Bank		Variable → Twitter Sentiment					Variable ← Twitter Sentiment				
	Lags	1	2	3	4	5	1	2	3	4	5
BMPS											
	r_t	-	-	-	-	-	-	-	-	-	-
	VV_t	-	-	-	-	-	-	-	-	-	-
	Vol_t	-	-	-	-	-	-	-	-	-	-
	CDS_t^{Sen}	*	-	-	-	-	-	-	-	-	-
	CDS_t^{Sub}	-	-	-	-	-	-	-	-	-	-
	$Bond_t^{Sub}$	-	-	-	-	-	-	-	-	-	-
ISP											
	r_t	-	-	-	-	-	-	-	-	-	-
	VV_t	-	-	*	*	-	-	-	-	-	-
	Vol_t	-	-	-	-	-	-	-	-	-	-
	CDS_t^{Sen}	-	-	-	-	-	-	-	-	-	-
	CDS_t^{Sub}	-	-	-	-	-	-	-	-	-	-
	$Bond_t^{Sub}$	-	-	-	-	-	-	-	-	-	-
UCG											
	r_t	-	-	-	-	-	-	-	-	-	-
	VV_t	*	*	-	-	-	-	-	-	-	-
	Vol_t	-	-	-	-	-	-	-	-	-	-
	CDS_t^{Sen}	-	-	-	-	-	-	-	-	-	-
	CDS_t^{Sub}	-	-	-	-	-	-	-	-	*	*
	$Bond_t^{Sub}$	-	*	-	-	-	-	-	-	-	-
DBK											
	r_t	-	-	-	-	-	-	-	-	-	-
	VV_t	*	-	-	-	-	-	-	-	-	-
	Vol_t	-	-	-	-	-	-	-	-	-	-
	CDS_t^{Sen}	*	-	-	-	-	-	-	-	-	-
	CDS_t^{Sub}	*	-	-	-	-	-	-	-	-	-
	$Bond_t^{Sub}$	*	*	-	-	-	-	-	-	-	-

TABLE 6.2: Granger Causality Tests for $Sent_{ts}$

Notes: The table depicts the Granger causality test results for the five banks included in our sample (BMPS, ISP, UCG, and DBK). For each bank we compute the test for the following financial indicators: i) stock returns (r_t), ii) volume (VV_t), iii) Volatility (Vol_t) computed as the daily range, iv) senior CDS spreads (CDS_t^{Sen}), v) subordinated CDS spreads (CDS_t^{Sub}), and vi) an average subordinated bond spread ($Bond_t^{Sub}$). On the left panel we test causality from financial variables to Twitter sentiment on that bank, while on the right panel we test causality from Twitter sentiment to the financial variable in each row. The Twitter sentiment TSI is computed as the simple average of sentiment on that bank. *, **, and *** indicate rejection of the null hypothesis of no Granger causality at 10%, 5%, and 1%, respectively.

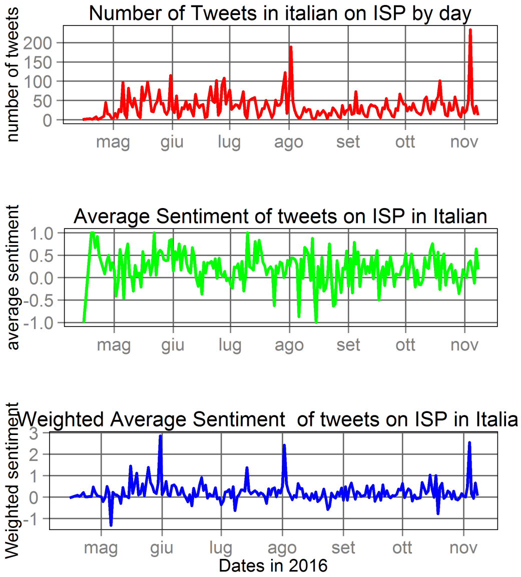


FIGURE 6.7: ISP Twitter data

Notes: The figure depicts the number of tweets in Italian for bank ISP on the top panel. The medium panel depicts the average sentiment of tweets on bank BMPS computed without any weightings on positive and negative tweets. The bottom panel depicts the weighted average sentiment of tweets on bank ISP where the weight is given by the daily number of tweets divided by the average number of tweets in the whole sample.

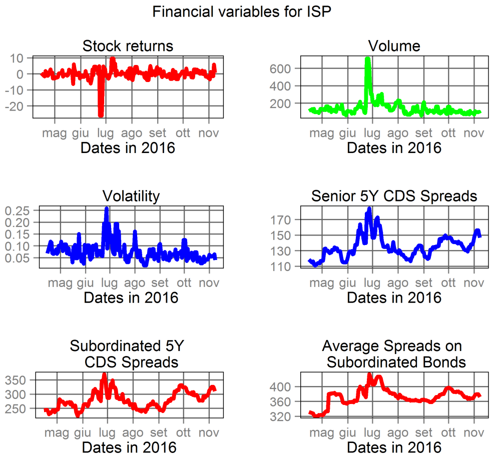


FIGURE 6.8: ISP financial data

Notes: The figure depicts for bank ISP from upper left panel to lower right panel: 1) stock returns, 2) volume of stocks traded, 3) volatility of stock returns, computed as the daily range (i.e. the difference between the highest and lowest price within each day), 4) the spread on the 5-year senior CDS, 5) the spread on the 5-year subordinated CDS, 6) the average spread on a selection of subordinated bonds issued by the bank.

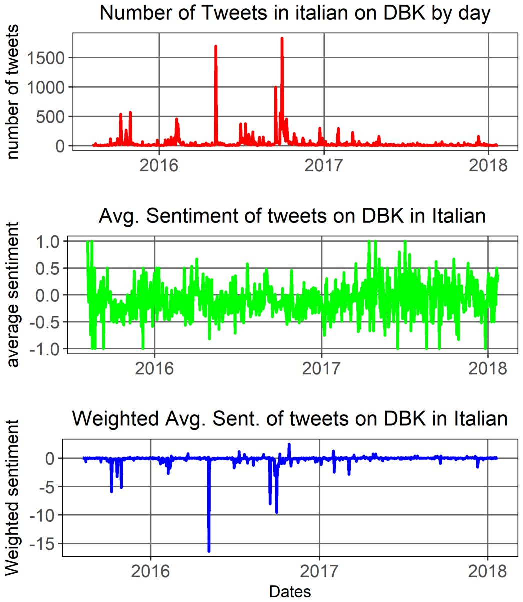


FIGURE 6.9: DBK Twitter data

Notes: The figure depicts the number of tweets in Italian for bank DBK on the top panel. The medium panel depicts the average sentiment of tweets on bank UCG computed without any weightings on positive and negative tweets. The bottom panel depicts the weighted average sentiment of tweets on bank DBK where the weight is given by the daily number of tweets divided by the average number of tweets in the whole sample.

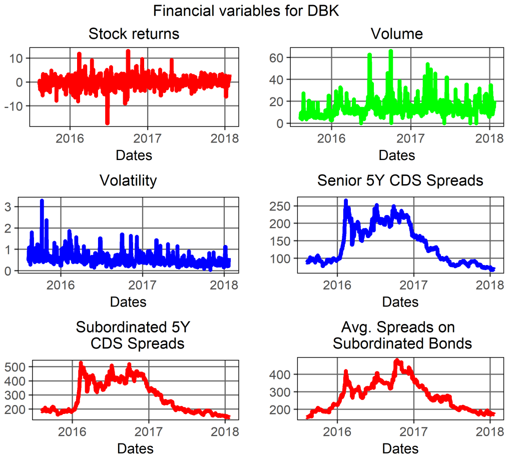


FIGURE 6.10: DBK financial data

Notes: The figure depicts for bank DBK from upper left panel to lower right panel: 1) stock returns, 2) volume of stocks traded, 3) volatility of stock returns, computed as the daily range (i.e. the difference between the highest and lowest price within each day), 4) the spread on the 5-year senior CDS, 5) the spread on the 5-year subordinated CDS, 6) the average spread on a selection of subordinated bonds issued by the bank.

Bank	**Ticker**	**Number of**	**Number of tweets**	**Number of retweets**	**Number of tweets**	**Average daily**
		total tweets	**in Italian**	**in Italian**	**used**	**# of tweets**
MPS bank	BMPS	783,150	606,006	345,510	260,496	341
Unicredit	UCG	27,435	23,400	4,407	18,993	25
Intesa S.Paolo	ISP	14,249	12,708	807	11,901	16
Deutsche Bank	DBK	2,422,559	79,593	45,394	34,199	45

TABLE 6.3: Descriptive statistics of tweets

Bank		Variable → Twitter Sentiment					Variable ← Twitter Sentiment				
	Lags	1	2	3	4	5	1	2	3	4	5
BMPS											
	r_t	-	-	-	-	-	***	***	***	***	***
	VV_t	-	-	-	-	-	***	***	***	***	***
	Vol_t	-	-	-	-	-	-	-	-	-	-
	CDS_t^{Sen}	-	-	-	-	-	*	-	-	*	**
	CDS_t^{Sub}	-	-	-	-	-	***	***	***	***	***
	$Bond_t^{Sub}$	-	-	*	**	**	-	-	-	-	-
ISP											
	r_t	-	-	-	*	*	-	-	-	-	-
	VV_t	*	*	*	*	*	-	-	-	-	-
	Vol_t	*	**	**	*	*	-	-	-	-	-
	CDS_t^{Sen}	-	-	-	-	-	-	-	-	-	-
	CDS_t^{Sub}	-	-	-	*	-	-	-	-	-	-
	$Bond_t^{Sub}$	-	-	-	-	-	**	**	**	**	*
UCG											
	r_t	-	-	-	-	-	-	-	-	-	-
	VV_t	-	-	-	-	-	-	-	-	-	-
	Vol_t	-	-	-	-	-	-	-	*	*	*
	CDS_t^{Sen}	*	**	**	**	**	-	**	**	*	*
	CDS_t^{Sub}	**	**	**	**	**	-	*	**	***	***
	$Bond_t^{Sub}$	*	**	**	**	**	-	-	-	-	-
DBK											
	r_t	-	-	-	-	-	*	**	**	***	***
	VV_t	-	-	-	-	-	-	-	***	***	***
	Vol_t	-	-	-	-	-	-	-	-	-	-
	CDS_t^{Sen}	*	*	-	*	*	-	-	-	-	-
	CDS_t^{Sub}	*	*	-	*	-	-	-	-	-	-
	$Bond_t^{Sub}$	-	-	-	-	-	-	-	-	-	***

TABLE 6.4: Granger Causality Tests for $Sent_{we}$

Notes: The table depicts the Granger causality test results for the five banks included in our sample (BMPS, ISP, UCG and DBK). For each bank we compute the test for the following financial indicators: i) stock returns (r_t), ii) volume (VV_t), iii) Volatility (Vol_t) computed as the daily range, iv) senior CDS spreads (CDS_t^{Sen}), v) subordinated CDS spreads (CDS_t^{Sub}), and vi) an average subordinated bond spread ($Bond_t^{Sub}$). On the left panel we test causality from financial variables to Twitter sentiment on that bank, while on the right panel we test causality from Twitter sentiment to the financial variable in each row. The Twitter sentiment TSI_we is computed as the weighted average of sentiment where the weights are the ratio between the number of tweets on that bank divided by the average number of daily tweets in the sample. *, **, and *** indicate rejection of the null hypothesis of no Granger causality at 10%, 5%, and 1%, respectively.

Chapter 7

Assessing banks' distress from news and numerical financial data

7.1 Summary

In this chapter we focus our attention on leveraging the information contained in financial news to enhance the performance of a bank distress classifier. The news information should be analyzed and inserted into the predictive model in the most efficient way and this task deals with the issues related to Natural Language interpretation and to the analysis of news media. Among the different models proposed for such purpose, we investigate a deep learning approach. The methodology is based on a distributed representation of textual data obtained from a model (Doc2Vec) that maps the documents and the words contained within a text onto a reduced latent semantic space. Afterwards, a second supervised feed forward fully connected neural network is trained combining news data distributed representations with standard financial figures in input. The goal of the model is to classify the corresponding banks in distressed or tranquil state. The final aim is to comprehend both the improvement of the predictive performance of the classifier and to assess the importance of news data in the classification process. This to understand if news data really bring useful information not contained in standard financial variables.

7.2 Introduction

Natural Language Processing (NLP), the interpretation of text by machines, is a complex task due to the richness of human language, its highly unstructured form and the ambiguity present at many levels, including the syntactic and semantic ones. From a computational point of view, processing language means dealing with sequential, highly variable and sparse symbolic data, with surface forms that cover the deeper structures of meaning.

Despite these difficulties, there are several methods available today that allow for the extraction of part of the information content present in texts. Some of these rely on hand crafted features, while others are highly data-driven and exploit statistical regularities in language. Moreover, once the textual information has been extracted, it is possible to enhance it with contextual information related to other sources different from text. The introduction of contextual information in the models is not always a straightforward process but requires a careful choice of the additional information provided in order to not increase noise by using irrelevant features. To accomplish such purpose, there are several methods of variable selection [Guyon and Elisseeff 2003] that can guide in the choice of the additional features for the model. The recent advancements in text analytics and the addition of contextual information aim at increasing the potential value of text as a source in data analysis with a special emphasis on financial applications (see for example [Nyman et al. 2015]). In this work, we focus on the issues of understanding and predicting banks distress, a research area where text data hold promising potential due to the frequency and information richness of financial news. Indeed, central banks are starting to recognize the usefulness of textual data in financial risk analytics [Bholat et al. 2015, Hokkanen et al. 2015].

If we focus only on the elicitation of information from textual data, we can find that among the statistical methods, many rely on word representations. Class based models, for example, learn classes of similar words based on distributional information, like Brown clustering [Brown et al. 1992] and Exchange clustering [Martin et al. 1998, Clark 2003]. Soft clustering methods, like Latent Semantic Analysis (LSA) [Landauer et al. 1998] and Latent Dirichlet Allocation [Blei et al. 2003], associate words to topics through a distribution over words of how likely each word is in each cluster. In the last years many contributions employ machine learning and semantic vector representations [Mikolov et al. 2013, Pennignton et al. 2014], lately using Long Short-Term Memory (LSTM) networks [Hochreiter and Schmidhuber 1997, Socher et al. 2013, Cho et al. 2014] to model complex and non-local relationships in the sequential symbolic input. Recursive Neural Tensor Networks (RNTN) for semantic compositionality [Socher et al. 2011, Socher et al. 2013] and also convolutional networks (CNN) for both sentiment analysis [Collobert et al. 2011] and sentence modelling [Kalchbrenner et al. 2014]. In this vein, [Mikolov 2012, Mikolov et al. 2013] and [Pennignton et al. 2014] have introduced unsupervised learning methods to create a dense multidimensional space where words are represented by vectors. The position of such vectors is related to their semantic meaning, further developing the work on word embeddings [Bengio et al. 2003] which grounds on the idea of distributed representations for symbols [Hinton et al. 1986]. The word embeddings are widely used in modern NLP since they allow for a dimensionality reduction compared to a traditional sparse vector space model. In [Le and Mikolov 2014], expanding the previous work on word embeddings, is presented a model capable of representing also sentences in a dense multidimensional space. Also in this case sentences are represented by vectors whose position is related to the semantic content of the sentence. In such a space sentences with similar semantic will be represented by vectors that are close to each other.

This recent rise of interest around text-based computational methods for measuring financial risk and distress is fuelling a rapidly growing literature. The most covered area is sentiment analysis to be correlated with events of interest. Many of the previous approaches have been based on hand-crafted dictionaries that despite requiring work to be adapted to single tasks can guarantee good results due to the direct link to human emotions and the capability of generalizing well through different datasets. Examples of this kind are the papers of [Nyman et al. 2015] and [Soo 2013]. The first analyses sentiment trends in news narratives in terms of excitement/anxiety and find increased consensus to reflect pre-crisis market exuberance, while the second correlates the sentiment extracted from news with the housing market. Despite the good results, there are applications where it could be preferable to avoid dictionaries in favour of more data driven methods, which have the advantage of higher data coverage and capability of going beyond single word sentiment expression. [Malo et al. 2014] provide an example of a more sophisticated supervised corpus-based approach, in which they apply a framework modelling financial sentiment expressions by a custom data set of annotated phrases.

Our contribution aims at demonstrating the feasibility and usefulness of the integration of textual and numerical data in a machine learning framework for financial predictions. Thus, the goal of the predictive model is to correctly classify stressed banks from both financial news and financial numerical data.

The rest of the chapter is organized as follows: in Section 7.3 we describe the machine learning framework, in Section 7.4 we illustrate the data and the predictive task, in Section 7.5 we present the experimental results with a sensitivity analysis on the network parameters and in Section 7.6 we discuss the conclusions of the work with hints on future developments.

7.3 Methodology

Machine learning systems benefit from their ability to learn abstract representations of data, inferring feature representations directly from data instead of relying on manual feature engineering. This capability is particularly exploited in deep learning models, which provides flexibility and potentially better performance [Schmidhuber 2015]. These characteristics are crucial in Natural Language Processing tasks where the ability to generalize across languages, domains and tasks enhances the applicability and robustness of text analysis. The framework applied in this chapter is an extension of the one developed in [Rönnqvist and Sarlin 2017] with the aim of predicting banks' distress from textual data. Their approach infers banks distress conditions from textual news using a machine learning system based on two steps:

- The first step comprises an unsupervised algorithm to compute the semantic vectors associated to a specific news text. Dense vector representations of sentences mentioning target banks are learned using the Distributed Memory Model of Paragraph Vectors (PV-DM) by [Le and Mikolov 2014] (here referred to as Doc2Vec). This algorithm represents each document by a dense vector which is trained to predict words appearing in the document. The semantic space obtained through algorithm has a lower dimensionality (600 in this case) compared to a Bag of Words representation and encodes the word semantics. In this representations in fact, words are represented by vectors whose distances reflect statistical properties of the language like synonymy, gender, verb tenses and many others. From this new space is easier to perform the classification task due to the reduced dimensionality and the wise positioning of the vectors that takes into account their semantic meaning.

- The second step of the framework performs a classification over the semantic vectors of sentences mentioning target banks through a supervised algorithm. The sentence representations are fed into a neural network classifier that is trained with distress event labels. The neural network architecture is constituted by an input layer with the same dimensionality of the semantic vectors (600 nodes), one hidden layer (50 nodes) and one output layer (2 nodes with stress prediction $e \in \{0,1\}$) that returns the tranquil or distressed status prediction.

In this work we modify the previous model of [Rönnqvist and Sarlin 2017] to integrate the financial numerical data and evaluate the performance gain obtained by their combination with news data. The financial data that we integrate contain information about bank accounting data, banking sector data and country macroeconomic data. In modifying the approach we kept the two-step structure of the previous framework. Thus, also in our case we previously compute the semantic representations of the textual data and then in a second step classify the bank status. Anyway, between the two steps we combine the semantic vectors with the numerical financial variables vectors. In this way, the classification performed in the second step takes into account both the information contained in the financial news and in the financial variables.

The approach used to learn the semantic vectors is a Distributed Memory Model Paragraph Vector [Le and Mikolov 2014]. In this model the semantic vector representation is learned by training a feed forward neural network to predict the words contained in a document by their word context (previous n and following n words) and a randomly initialized semantic vector (sentence ID). The word contexts, used as features to predict the target words are fixed-length and sampled from a sliding window over the sentence. While training the network, the semantic vector gets updated by the training algorithm so that its representation positively contributes in predicting the next word and thus works as a semantic representation of the entire sentence (or text sequence). In this way the sentence vector works as a memory for the model that once trained captures the semantics of continuous sequences. The sentence ID, in fact, can be thought of as an extra word representing the sentence as a global context on which the prediction of the next word is conditioned. Despite the random initialization of the semantic vectors, they gradually improve the capability of capturing the semantic of the sentence during the training. The training is performed by stochastic gradient descent with the gradients computed by the backpropagation algorithm. Formally, the training procedure seeks to maximize the average log probability:

$$\frac{1}{t+n}\sum_{i=1}^{t-n}\log p(w_{i+n+1}|s,w_i,...,w_{i+n}) \tag{7.1}$$

over the sequence of training words $w_1, w_2, ..., w_t$ in sentence s with word context of size n.

After being trained, the semantic vectors can be used as features for representing the sentence information content (e.g., in place of its Bag of Words representation). These features can be feed directly to conventional machine learning techniques such as logistic regression, support vector machines, neural networks or K-means clustering.

The algorithm can be used both to compute the semantic vectors of the sentences on which is trained and also to infer the semantic vectors of new unseen sentences. In the first case the model learns the semantic vectors along with the word vectors from the training via backpropagation and gradient descent by minimizing the word prediction error on the training corpus. In the second case the semantic vectors are calculated by gradient descent and backpropagation while keeping the word vectors and the other model parameters fixed.

As first step we compute the semantic representations of the sentences mentioning banks in our corpus. To obtain valid sentence representations for specific domains it's important to train the model on large enough corpora that also contain task-specific texts. In our case we would like our model to capture both the the general properties of the English language and the context specific terms and expressions related to banks. To do so, we run the model on the entire corpus of ca. 262,000 articles that we have disabling the sentence ID vector for those sentences that don't contain any bank occurrences. In this way the word representations that the model internally builds can take advantage of a larger quantity of text. The dimensionality of the semantic vectors (600) and the word context size of the algorithm (5) have been optimized by cross-validation.

The second step performs the classification task on the combination of the financial news and financial numerical information. It receives in input the news textual data on the banks, in form of sentence level semantic vectors V_s, and the vector of numerical financial data F_s for the corresponding bank in corresponding period. The classification model is a three layers fully connected feed forward neural network. The neural network has an input layer with 612 nodes, 600 input nodes for the semantic vector V_s and 12 input nodes for the numerical data F_s. After the input layer it has a 50 nodes hidden layer and a 2 nodes output layer with softmax activations $e \in \{0,1\}$ to encode the distress or tranquil status (see Figure 7.1). The network is trained by Nesterov's Accelerated Gradient Descent [Nesterov 1983] to minimize the cross-entropy loss function. Hence, the objective is to maximize the average log probability:

$$\frac{1}{|S|}\sum_{s\in S}\log p(e_s|V_s,F_s) \tag{7.2}$$

In the trained network, the posterior probability $p\left(e_s = 1|V_s,F_s\right)$ reflects the relevance of sentence s and the corresponding financial variables to the modelled event type.

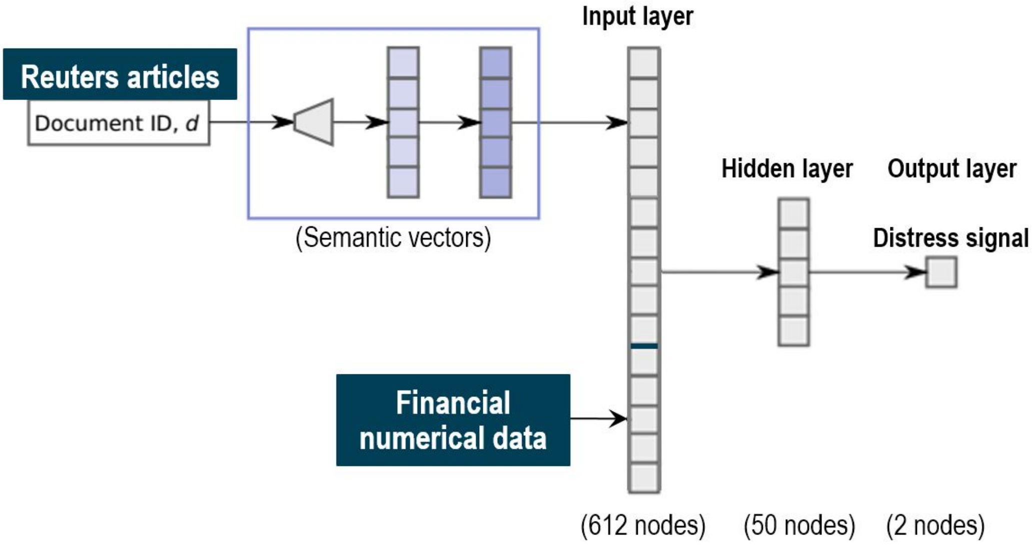


FIGURE 7.1: Structure of the model

7.4 Data

As described in the previous section, we leverage two types of data. We rely on textual and numerical data, aligned together by time and entities (banks), to classify bank distressed or tranquil conditions. The distress events dataset contains information on dates and names of the involved entities, relating to the specific type of distress event to be modelled. The textual and financial numerical datasets accordingly contain respectively bank related articles and financial figures. The textual and numerical data are aligned and linked to a particular event matching the date and occurrences of the entity name within the sentence. For the financial numerical data with quarterly frequency the date of the event is matched with the corresponding quarter. The model is then trained in a supervised framework to associate specific language and financial figures with the target bank status for that period (tranquil or distressed).

7.4.1 Textual and numerical data

The textual data object of the study are part of a news articles database from Reuters online archive spanning the years from 2007-Q1 to 2014-Q3. The original data set includes 6.6M articles, for a total of ca. 3.4B words. In order to select only articles related to the considered banks, we have looked at bank names occurrences and selected only those articles with at least one occurrence. Bank name occurrences are located using a set of patterns defined as regular expressions that cover common spelling variations and abbreviations of the bank names. The regular expressions have been iteratively developed on the data to increase accuracy, with a particular attention on avoiding false positives. As a result from the entire corpus we retrieve ca. 262,000 articles mentioning any of the 101 target banks. Successively the articles are split into sentences and only the sentences with bank name occurrences are kept. We integrate financial contextual information through a database of distress related indicators for banks. The numerical dataset is composed of 12 variables for 101 banks over the period 2007-2014 with quarterly frequency. In Table 7.1 we list the considered numerical variables: among them there are information on bank-level balance sheet and income statement data, as well as country-level banking sector and macro-financial data.

The three bank-specific variables are the ratio of tangible equity to total assets, the ratio of interest expenses to total liabilities and the NPL reserves to total assets ratio. The three banking-sector features are the mortgages to loans ratio (4-months change), the ratio of issued debt securities to total liabilities (4-months change) and the ratio of financial assets to GDP. The six Macro financial level features are the House price gap (Deviation from trend of the real residential property price index filtered with the Hodrick-Prescott Filter [Hodrick and Prescott 1980] with a smoothing parameter λ of 1600), the international investment position from the ECB Macroeconomic Imbalance Procedure (MIP) Scoreboard, the country private debt, the government bond yield (4-months change), the credit to GDP ratio and the credit to GDP 1-yr change.

Bank Level	**Bank Sector Level**	**Macro Level**
Capital to asset	Mortgages to loans	House price gap (Deviation from trend of the real residential property price index)
Interest to liabilities	Securities to liabilities d4	Macroeconomic Imbalance Procedure (MIP), international investment position
Reserves to asset	Financial assets to gdp	Private debt
-	-	Government bond yield
-	-	Credit to gdp
-	-	Credit to gdp delta over 12 months

TABLE 7.1: List of available numerical variables.

In Table 7.2 we report summary statistics of the analyzed numerical variables.

Variable	**Mean**	**Variance**	**Standard Deviation**	**Kurtosis**
Capital to asset	2.5	10.2	3.2	21.5
Reserves to asset	4.2	8.5	2.9	4.3
Interest to liab	3.4	8.5	2.9	104.6
Financial assets to gdp	385.0	134,365.2	366.6	33.2
Mortgages to loans d4	0.2	1.7	1.3	0.1
Securities to liab d4	-12.0	1,342,234.9	1,158.5	105.7
Credit to gdp	140.2	2,623.4	51.2	0.0
Credit to gdp d12	13.7	479.1	21.9	0.5
House Price Index rt16 gap	-2.5	33.7	5.8	6.8
International Investment Position	-21.0	2,967.7	54.5	0.0
Private Debt	188.2	4938.7	70.3	0.1
Gov Bold Yield d4	0.0	11.4	3.4	23.5

TABLE 7.2: Summary statistics of available numerical variables.

Then, we match the distress events with the available textual news data. The events comprehends bankruptcies, direct defaults, government aid and distressed mergers as presented in [Betz et al. 2003]. The distress events in this dataset are of three types. The first type of events include bankruptcies, liquidations and defaults, with the aim of capturing direct bank failures. The second type of events comprises the use of state support to identify banks in distress. The third type of events consists of forced mergers, which capture private sector solutions to bank distress. The inclusion of state interventions and forced mergers is important to better represent bank distress since there have been few European direct bank failures in the considered period. Bankruptcies occur if a bank net worth falls below the country-specific guidelines, whereas liquidations occur if a bank is sold and the shareholders do not receive full payment for their ownership. Defaults occur if a bank failed to pay interest or principal on at least one financial obligation beyond any grace period specified by the terms or if a bank completes a distressed exchange. The distress events are formally considered to start when a failure is announced and end at the time of the 'de facto' failure.

A capital injection by the state or participation in asset relief programs (i.e., asset protection or asset guarantees) is an indication of bank distress. From this 'indicator' are excluded liquidity support and guarantees on banks' liabilities since they are not used for defining distressed banks. The starting dates of the events refer to the announcement of the state aid and the end date to the execution of the state support program. Distressed mergers are defined to occur if (i) a parent receives state aid within 12 months after a merger or (ii) if a merged entity exhibits a negative coverage ratio within 12 months before the merger. The dates for these two types of distress events are defined as follows, respectively: (i) the starting date is when the merger occurs and the end date when the parent bank receives state aid, and (ii) the start date is when the coverage ratio falls below 0 (within 12 months before the merger) and the end date when the merger occurs. Thus far, data at hand assign a unique label for the stress events, not allowing a more detailed descriptive summary of the three event types.

7.4.2 Data Integration

The following step in the data preparation has been the addition of numerical financial data to the text news database. The numerical data were aligned with the set of sentences in which bank names occurred. The purpose was to match each and every mention of a bank with the corresponding numerical financial data aligned according to the same time horizon. Since the news and the financial data have different frequency, in particular, news have higher frequency while financial data are reported quarterly, the latter are replicated several times to match with the former. For each news regarding a bank within a given quarter, financial data are replicated and appended to the semantic vector of the news. The alignment between numerical and textual data resulted in the removal of some banks from the dataset due to missing data, causing a reduction from 101 to 62 target institutes (Table 7.3) and from about 601,000 to 380,000 news sentences. After cleaning the dataset, numerical data have been normalized with a standard approach by subtracting the mean value from each numerical variable of the dataset and dividing it by the standard deviation. The resulting input vector for the 612 dimensional input layer of the neural classifier receives in input a 612 dimensional vector obtained from the concatenation of the 600-dimensional semantic vector coming from the unsupervised modelling, described in Section 7.3, with the 12-dimensional numerical financial data vector. The dataset is then split into five folds, three for training, one for validation and one for testing according to a cross-validation scheme. The folds are created so that all the data regarding a given bank are in the same fold. The framework we apply is composed of an unsupervised algorithm and a supervised neural network classifier. To train the classifier a label indicating the distressed or tranquil status of the bank is provided. The dataset has been labelled according to the bank status with 0 indicating tranquil and 1 distressed. The proportions of the two classes are highly unbalanced: 93% of the data-points are associated to a tranquil status and only the remaining 7% are associated to distress events. Such imbalance of the classes has a significant impact both on the training and on the evaluation of the model. Regarding the training, it is important that the model is able to generalize also from the few distress examples, while for the evaluation it could be useful to include other performance measures that the accuracy. A trivial model that always predicts the tranquil status would achieve a 93% accuracy, thus it would be interesting to measure the improvements against this baseline. Moreover the user is likely interested in weighting differently first error and second error types, especially if we consider potential early warning applications of this model. The usefulness measure, introduced in [Sarlin 2013], satisfies these requirements.

Financial Institution	Country	Financial Institution	Country	Financial Institution	Country
Aareal Bank	DE	Carnegie Investment Bank	SE	Kommunalkredit	AT
ABN Amro	NL	Commerzbank	DE	LBBW	DE
Agricultural Bank of Greece	GR	Credit Mutuel	FR	Lloyds TSB	UK
Allied Irish Banks	IE	Credito Valtellinese	IT	Max Bank	DK
Alpha Bank	GR	Cyprus Popular	CY	Monte dei Paschi di Siena	IT
Amagerbanken	DK	Danske Bank	DK	National Bank of Greece	GR
ATE Bank	GR	Dexia	FR	Nordea	SE
Attica Bank	GR	EBH	DK	NordLB	DE
Banca Popolare di Milano	IT	EFG Eurobank	GR	Nova ljubljanska banka Group (NLB)	SI
Banco Popolare	IT	Erste Bank	HU	OTP Bank Nyrt	HU
Bank of Cyprus Public Co Ltd	CY	Fionia (Nova Bank)	DK	Piraeus Bank	GR, CY
Bank of Ireland	IE	Fortis Bank	LU, NL, BE	Pronton Bank	GR
Banque Populaire	FR	HBOS	UK	RBS	UK
Bawag	AT	Hellenic	GR	Roskilde Bank	DK
BayernLB	DE	HSH Nordbank	DE	Societe Generale	FR
BBK	ES	Hypo Real Estate	DE	Swedbank	SE
BNP Paribas	FR	Hypo Tirol Bank	AT	T-Bank	GR
BPCE	FR	IKB	DE	UNNIM	ES
Caixa General de Depositos	PT	ING	NL	Vestjysk	DK
Caja Castilla-La Mancha	ES	Irish Nationwide Building Society	IE		
CAM	ES	KBC	BE		

TABLE 7.3: List of considered financial institutions

7.5 Results

The experimental results confirm that the integration of numerical and textual data amplifies the prediction capability of the model compared to the inclusion of only textual data. The distress events in the database represent only 7% of the cases, resulting in very skewed training classes as explained earlier. Moreover, given the nature of the problem, the identification of distress situations, it could be useful to weight differently false positives and false negatives. In an early warning application, a sensitive system is often preferable since a further investigation phase follows the detection of a warning. These peculiarities have to be taken into account during the evaluation of the model.

7.5.1 Evaluation and experimental results

For the evaluation of our model we resort to the relative usefulness as measure of performance. The relative usefulness (U_r), introduced in [Sarlin 2013] is a measure that allows to set the error type preference (μ) and to measure the relative performance gain of the model over the baseline compared to the performance gain over the baseline of a perfect model. The index is computed starting from the probabilities of the true positive (TP), false positive (FP), true negative (TN) and false negative (FN). With these we can define the model loss L_m (Equation 7.4) and a baseline loss L_b set to be the best guess according to prior probabilities $p(obs)$ and error preferences μ (Equation 7.3).

$$L_b = min \begin{cases} \mu * p(obs = 1) \\ (1-\mu) * p(obs = 0) \end{cases} \tag{7.3}$$

$$L_m = \mu * p(FN) + (1-\mu) * p(FP) \tag{7.4}$$

The absolute Usefulness (U_a) and the relative Usefulness (U_r) are directly derived from the loss functions:

$$U_r = \frac{U_a}{L_b} = \frac{L_b - L_m}{L_b} \tag{7.5}$$

The absolute Usefulness U_a of a model corresponds to the loss "generated" by the model subtracted from the loss of ignoring it L_b. From Equation 7.5 we can see that the relative usefulness is equal to 1 when the model loss (L_m) is equal to 0, thus when the model is a perfect classifier. As a consequence, the relative usefulness measures the gain over the baseline compared to the gain that an ideal model would achieve. U_r reports U_a as a percentage of the Usefulness that one would gain with a perfectly performing model. This measure highlights the fact that achieving well-performing, useful models on highly imbalanced data is a difficult task. To compute the relative usefulness (U_r) we have set the error type preference (μ) equal to 0.9 in accordance with the indications of previous studies like [Betz et al. 2003] and [Constantin et al. 2016] on the importance of signalling every possible crisis at cost of some false positive (FP) (setting μ = 0.9 we are implying that missing a crisis is about 9 times worse than falsely signalling one). This is especially true if following the warning signal, a further investigation action is triggered. To evaluate distress condition of a bank over a period, the predictions are aggregated on a monthly basis by bank entity. This is done by averaging the predictions at the single sentence level by month for each different bank. This has been done to take into account the textual information available over the past month period. As a result of this procedure, the classification task can be summarized as understanding which banks are in distress status month by month based on the news sentences and numerical data available over the previous month.

To evaluate the model on this classification task, we have trained it fifty times on the same dataset, recording the relative usefulness (U_r) result after each run and then averaging them. For each of the fifty trainings, the folds are resampled and the neural net is randomly initialized. To quantify the gain obtained from merging numerical and textual data we have done three different experiments, training the model respectively with textual data only (Figure 7.2, left), numerical data only (Figure 7.2, center) and numerical and textual data together (Figure 7.2, right). As it is possible to see from Figure 7.2 the case with textual data alone achieves an average relative usefulness of 13.0%, while the case with numerical data alone shows an average relative usefulness of 31.1%. The combination of these two dataset and their exploitation in the model grants an average relative usefulness of 43.2%, thus it positively enhances the prediction capability of the model. From these results we can also understand that, as expected, the financial numerical data hold the majority of the informative potential necessary for the labelling task but that the addition of textual information provides a non-negligible 12.1% improvement to the relative usefulness of the model.

7.5.2 Classifier tuning

We have run a sensitivity analysis exploring different neural network configurations while training it with the Nesterov Accelerated Gradient Descent algorithm from [Nesterov 1983]. We have tested different hidden layers sizes, numbers of layers, learning rates, regularization parameters and dropout fractions [Hinton et al. 2012]. For choosing the final network configuration, we applied the Occam's razor principle always preferring the simpler structure able to achieve a given performance. Thus, where performance is not reduced excessively, we try to select the network structure with fewer layers and fewer hidden nodes; this also helps to have better generalization and reduce overfitting. In terms of hidden layers number, the network with one hidden layer (three layers in total including input and output) performs slightly better than those with more layers. We tested up to three hidden layers (5 layers in total, including input and output layers) and verified that the performance was monotonically

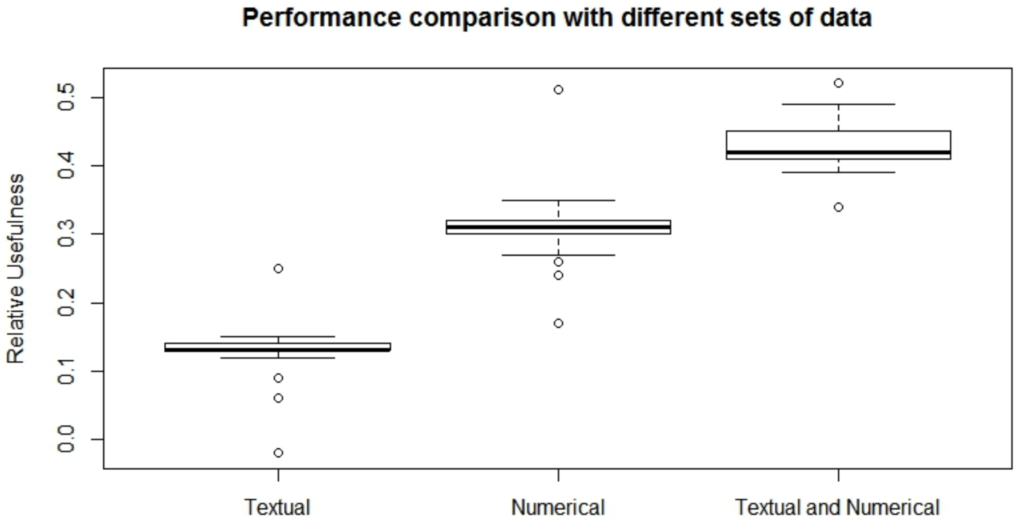


FIGURE 7.2: Comparison of the relative usefulness obtained with the textual financial data (left), numerical financial dataset (center) and with their combination (right)

decreasing. Regarding the number of hidden nodes, the network configuration that achieved the best relative Usefulness had 50 hidden nodes with a learning rate α of $5e$-4 combined with an L_1 regularization parameter λ of $1e$-5. The parameter that mostly affects the results is the number of nodes in the hidden layer. The results of the sensitivity analysis on the hidden nodes number (with regards to one hidden layer network configuration) are reported in Figures 7.3, 7.4, 7.5 respectively for the case including textual data alone, financial numerical data alone and the combination of the two. The range of hidden nodes in the three sensitivities is different because the input vectors in the three cases have very different dimensions, 600 input nodes when considering only textual data, 12 input nodes when considering only numerical data and 612 input nodes when including both numerical and textual data. We do not investigate extensively the textual data case which has already been studied in [Rönnqvist and Sarlin 2017]. Regarding the numerical based case, we can notice that we have a range of hidden layer size comprised between 10 and 20 nodes where performances are stable and the relative Usefulness is around 30%. For the combined input (Numerical and Textual) we expected the right number of hidden nodes to be similar to the Textual data case since the input dimensionality is similar (600 and 612). In fact, we can see that there is a range around 50-60 hidden nodes where performance is stable around a relative Usefulness of 40%.

7.6 Conclusions

In this chapter we have presented an approach for the integration of financial numerical data and financial news data into a single machine learning framework. The aim is to improve performances of bank distress conditions identification through the combination of these two data sources. The implemented model processes textual data through an unsupervised neural network model, Doc2Vec, converting the documents sentences into sentence vectors. The derived sentence vectors are then concatenated with the financial numerical data to form a single input vector. Each of these vectors becomes the input to a supervised classifier, a three layers fully connected neural network.

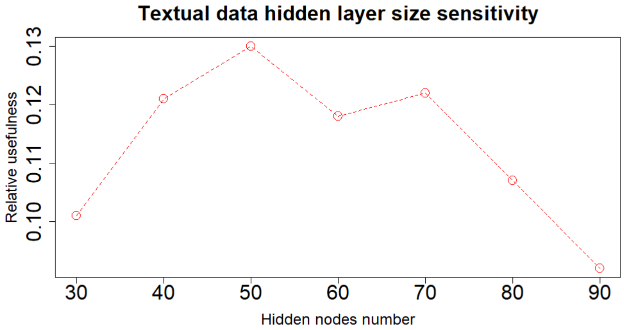


FIGURE 7.3: Textual data - sensitivity analysis on the number of nodes of the hidden layer

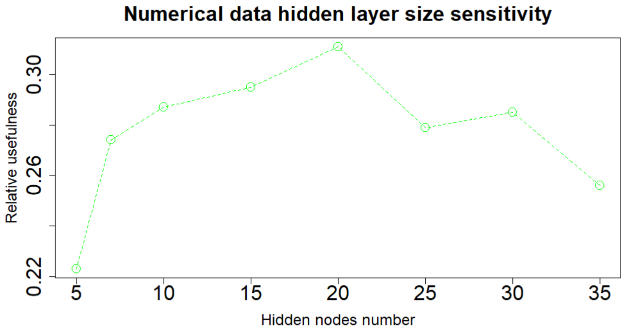


FIGURE 7.4: Numerical data - sensitivity analysis on the number of nodes of the hidden layer

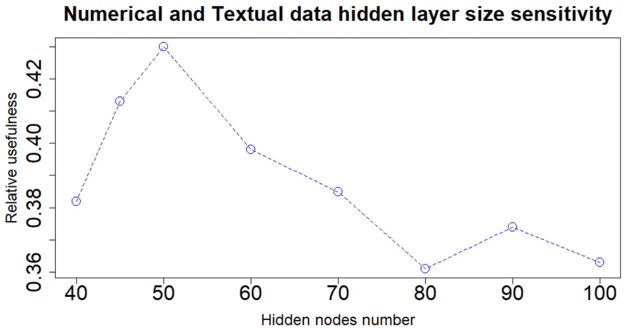


FIGURE 7.5: Numerical and Textual data - sensitivity analysis on the number of nodes of the hidden layer

The classification task was characterized by a high data imbalance among the classes which poses concerns for both the model training and evaluation. The implemented model has shown to be able to learn the combinations of banks' financial conditions and news semantic content that are more frequently associated with distress conditions. This is reflected in the improved performance obtained when including both news data and financial numerical data as input to the model. In this case in fact, the model achieves an average relative usefulness of 43.2%, compared to 31.1% when using only numerical data and 13.0% when utilizing only textual data. The sensitivity analysis performed on the model supports these results indicating stability within a certain range of architectures.

Some limitations of this model reside in the way the news are processed and converted into vectors and how they are fed to the network to classify the distress. Methods like Doc2Vec with Distributed Memory approach in fact, while not using a pure bag of word approach, still ignore important text information to truly understand a sentence and not only its topic or its average sentiment. For example, long range dependencies in the text are not considered and polysemous words and mixed word polarities can also affect the performance of this algorithm. Moreover, Doc2Vec performs significantly better when trained on a large quantity of text similar to the application domain. This quantity of texts was available in our study but could pose a limit to applications in niche specific domains or its extension to less widespread languages. In the last years there have been many improvements in the NLP field that can help overcome these limitations. A particularly interesting class of models are the so-called Sequence to Sequence RNNs, that recently have become very popular. These models are composed of two RNNs (one encoder and a decoder) that are trained in an unsupervised setting to reconstruct their own input text. Sequence to Sequence architectures pre-trained on financial and bank related text could be used as a substitute for the Doc2Vec representation in our approach. Differently from Doc2Vec these models consider explicitly the word order and long range dependencies over the entire text input sequence. As a result they can provide more accurate text vector representations. Furthermore, it is possible to augment the model capability providing few additional manually engineered features like a gazette of words with positive/negative polarity from a financial stability point of view.

An additional future work direction that could improve this framework as an early warning tool would be considering the news dynamic evolution. In this work, news are aggregated at monthly level, thus sub-monthly dynamics are lost. Using a RNN as distress classifier, it would be possible to sequentially feed the news vectors into the network. In this way considering daily or weekly news aggregation it would be possible to take into account also these dynamic effects (e.g. overall negative sentiment but with a positive trend in the last weeks).

The methodology here applied is general and extensible to other problems were the integration of text and numerical covariates can improve classification and early warning performances. Similarly interesting results are to be expected in areas where textual data hold information with higher granularity and frequency, directly influencing the data to be predicted in the short run like in the case of financial markets.

Chapter 8

SentITA a python sentiment analysis tool for Italian

8.1 Summary

This chapter describes the SentITA system a sentiment analysis tool for Italian. The system, based on Deep Learning, is focused on general domain sentiment analysis at sentence level. The underlying model is a Bidirectional Long Short Term Memory network with attention that exploits word embeddings and sentiment specific polarity embeddings. The model has been trained with a custom dataset of sentence polarities in Italian. The dataset has been created by combining together labelled sentences from different sources. In particular, two Italian sentiment analysis challenges (Sentipolc2016 and Absita2018) and manually labelled sentences, for a total of ca. 14,000 labelled sentences. The model also leverages grammatical information from POS tagging and NER tagging. The system participated in both the Aspect Category Detection (ACD) and Aspect Category Polarity (ACP) tasks of the ABSITA2018 challenge achieving the 5^{th} place in the ACD task and the 2^{nd} in the ACD task. In an attempt to reduce the gap between sentiment analysis tools in English and in Italian, and ease future researches that leverage sentiment analysis, a python package implementing the model and the relative code has been publicly released and a brief guide on its installation is included in the chapter.

8.2 Introduction

Sentiment analysis is a task of Natural Language Processing (NLP) that investigates people's opinions towards different matters: products, events, organisations [Bing 2012]. Sentiment analysis adoption has been growing constantly in the last years with the rapid and wide diffusion of social networks, microblogging applications and forums. This media in fact, have made possible to gather a huge volume of user opinions, sentiments, emotions, appraisals, and attitudes towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. Mining the shear amount of opinions and informations passing through these services, if carefully done, can be of both social and commercial interest. Opinions in fact, are central to many human activities and are key influencers of our behaviours. Our beliefs and the choices we make are often conditioned upon how others see and evaluate the world. In fact, in many decision both individuals and organizations seek out the others' opinions. Lately many researches have given evidences that by analysing sentiment of social-media content it might be possible to predict some economic and social phenomena like the size of the markets [Bollen et al. 2011] or unemployment rates over time [Antenucci et al. 2014] or events like movies' box office performance and general elections [Heredia et al. 2016].

Sentiment analysis is far from being a solved problem of NLP and it has many factors that make it difficult compared, for example, to text topic classification.

We could try initially to identify the polarity of opinions by using a set of keywords; this is a simple but already effective approach in topic classification. Unfortunately, the results of an early study [Pang et al. 2002] on movie reviews show that identifying the right set of keywords is not so easy and doesn't achieve the best performances. In their study in fact, the use of the subjects' lists of keywords achieves only about 60% accuracy when employed within a straightforward classification policy. Word lists of the same size but chosen based on examination of the corpus' statistics perform significantly better, achieving almost 70% accuracy.

Data-driven approaches like machine learning techniques based on unigram models can achieve still better accuracies, over 80% [Pang et al. 2002], much higher than the performance based on hand-picked keywords. However, this level of accuracy is generally lower than the performance one would expect in typical topic-based binary classification. In fact, compared to topics, sentiment it is often expressed in a more subtle manner, making it difficult to be identified by any sentence or document's terms when considered in isolation. Even strong opinions are not always easy to recognize because in many cases is not possible to identify them from specific keywords or phrases in the sentence but rather from a combination of words given a certain context.

Moreover, even if the general notion of positive and negative opinions is fairly consistent across different domains, sentiment and subjectivity are quite context-sensitive and domain dependent. The same expression can be associated with opposite sentiment in different domains. For example, "go read the book" indicates positive sentiment for book reviews, but negative sentiment in the context of movie reviews.

Furthermore, it has great importance also modelling the discourse structure. While the overall topic of a document can be guessed by the text content regardless of the order in which different subjects are presented (e.g. Bag of Words representation), for opinions different orders can result in a completely opposite overall sentiment polarity. For sentiment analysis order effects can completely overwhelm frequency effects and in general, modelling sequential information and discourse structure is more crucial than for topic-based text categorization [Pang et al. 2008]

In the struggle to overcome these challenges researchers has developed numerous techniques for various sentiment analysis tasks. These techniques include both unsupervised and supervised methods. Among the unsupervised methods many exploit sentiment lexicons, grammatical analysis and syntactic patterns. In the supervised setting, most of the supervised machine learning methods have been tested (Support Vector Machines (SVMs), Logistic Regression, Maximum Entropy, Naïve Bayes, etc.) with different feature combinations [Liu 2015].

Recently in the last ten years, deep learning has emerged as a powerful machine learning technique achieving state-of-the-art results in many application domains, ranging from computer vision to speech recognition to NLP. Sentiment analysis makes no exception and also in this task the application of deep learning has pushed forward the state of the art. Among the deep learning frameworks applied to sentiment analysis, many employ a combination of semantic vector representations [Mikolov et al. 2013, Pennignton et al. 2014] and different deep learning architectures. Long Short-Term Memory (LSTM) networks [Hochreiter and Schmidhuber 1997, Socher et al. 2013, Cho et al. 2014] have been applied to model complex and long term non-local relationships in both word level and character level text sequences. Recursive Neural Tensor Networks (RNTN) have shown great results for semantic compositionality [Socher et al. 2011, Socher et al. 2013] and also Convolutional Neural Networks (CNNs) for both sentiment analysis [Collobert et al. 2011] and sentence modelling [Kalchbrenner et al. 2014] have performed better than previous state of the art methodologies. All these methods in most of the applications receive in input a vector representation of words called word embeddings. [Mikolov 2012, Mikolov et al. 2013] and [Pennignton et al. 2014], further expanding the work on word embeddings from [Bengio et al. 2003], that grounds on the idea of distributed representations for symbols [Hinton et al. 1986], have introduced unsupervised learning methods to create dense multidimensional spaces where words are represented by vectors. The position of such vectors is related to their semantic meaning and grammatical properties. [Le and Mikolov 2014] continuing on this research direction, develops also a model capable of representing sentences and documents in a dense multidimensional space. In this case too, sentences are represented by vectors whose position is related to their semantic content. Also in this space representation similar sentences are represented by vectors that are close to each other.

Word embeddings currently are widely used in most of the NLP tasks. They allow for a dimensionality reduction compared to traditional sparse Vectors Space Models (VSMs) [Salton 1975] and they are often used as pre-trained initialization for the first embedding layers of the neural networks in NLP tasks. In fact, word embeddings have been the core methodology for transfer learning for most of NLP tasks in the last years. They allow to relieve the network from the burden of learning the word semantics and how words relate to each other in text. Normally, in transfer learning applications, the first embedding layer of the NLP neural networks is initialized with a word embeddings weight matrix that is pre-trained with unsupervised methods on huge corpora. In this way the model, instead of being initialized randomly, has already learned (by transfer) a wiser word representation that encodes part of the language statistical regularities like word semantic, gender, plurality, verb tenses and many others.

Recently in 2017-2018 there has been a lot of interest around unsupervised or semi-supervised transfer learning methodologies for NLP that try to improve on word embeddings. The research is focused on algorithms that provide more than just pre-trained vectors of words, providing pre-trained vectors for sentences or blocks of sentences. Two methods have obtained promising results [kiros et al. 2015] and [Howard et al. 2018]. The first proposes an unsupervised learning of a generic, distributed sentence encoder. Using the continuity of text from books, the authors train an encoder-decoder model that tries to reconstruct the surrounding sentences of an encoded passage. Sentences that share semantic and syntactic properties are thus mapped to similar vector representations (in analogy with the word vectors). In the second, the authors propose a transfer learning method, based on a Universal Language Model Fine-tuning (ULMFiT), that can be applied to different NLP tasks, and they also introduce several key techniques for language models fine-tuning. This method too, produces a vector representation of sentences or text sequences in general. These approaches allow to transfer from one NLP task to another also the capability of modelling the discourse structure learnt by the model. In this way the network is relieved not only from the burden of learning word representations but also from that of learning to model the language trough sentence representations. Like word embeddings these methodologies are very interesting because they are unsupervised and can be trained on extremely vast unlabelled corpora. This allows to reduce the amount of supervised training required and to develop models with a limited number of labelled examples. This is of paramount importance since it makes many NLP tasks practical also for languages with fewer labelled resources available like Italian. These techniques excel in long text sequences because they take advantage of long term dependencies of text (what has been said in the previous sentences), specially ULMfiT, and thus are particularly suitable for topic classification and sentiment analysis of longer texts (like the IMDB dataset).

When working with isolated and short sentences, often with a specific writing style, like tweets or phrases extracted from internet reviews many long term text dependencies are lost and not exploitable. In this situation it is important that the model learns both to pay attention to specific words that have key roles in determining the sentence polarity like negations, magnifiers, adjectives and to model the discourse but with less focus on long term dependencies (due to the text brevity). For this reason, deep learning word embedding based models augmented with task specific gazettes (dictionaries) and features, represent a solid baseline when working with these kind of datasets [Nakov et al. 2016, Attardi et al. 2016, Castellucci et al. 2016, Cimino et al. 2016, Deriu et al. 2016].

In this chapter we present a word embedding based model, augmented with several additional features, for sentiment analysis on short Italian sentences and reviews. In the system in fact, a polarity dictionary for Italian has been included as input feature to the model. Moreover, every sentence during preprocessing is augmented with its NER tags and POS tags which then are fed as input to the model. Thanks to the inclusion of these relevant features in combination with word embeddings and an attentional bidirectional LSTM recurrent neural network architecture, the model already achieves useful results with some thousands labelled examples.

The remainder of the chapter presents the model, the experiments on the ABSITA 2018 task and the SentITA package installation and usage guide. In Section 8.3 the model architecture is described; in Section 8.4 we explore the data used to train the model; in Section 8.5 the model training and its performances are discussed along with a brief guide on the installation and usage of SentITA in Subsection 8.5.3; finally in Section 8.6 the conclusions of this work with the next improvement steps of the system are discussed.

8.3 Methodology

The implemented model is an Attentional Bidirectional Recurrent Neural Network with LSTM cells. It operates at words level and therefore each input sentence is represented as a sequence of words representations in the form of vectors. These vectors are sequentially fed to the model one after another until the sentence word sequence has been entirely used up. In this setup, one sentence sequence matched with its polarity scores represent a single labelled data point for the model. [1]

The input to the model are sentences up to 35 words of length, with shorter sentences left-padded with zero values to this length and longer sentences cut to this length. However, it is possible to apply the model also to longer texts by splitting them in sentences, calculating each sentence polarity separately and then aggregating the results at document or paragraph level. Each word of the input sentence sequence is represented by five vectors corresponding to 5 different features that are: high dimensional word embeddings, word polarity, word NER tag, word POS tag, custom low dimensional word embeddings. The high dimensional word embeddings are the pre-trained Fastext embeddings for Italian [Grave et al. 2018]. They are 300-dimensional vectors computed using the skip-gram model described in [Bojanowski et al. 2016] with default parameters. The word polarity is obtained from the OpeNER Sentiment Lexicon Italian [Russo et al. 2016]. This freely available Italian Sentiment Lexicon contains a total of 24,293 lexical entries annotated for positive/negative/neutral polarity. It was semi-automatically developed using a propagation algorithm starting from a list of seed keywords and manually reviewing the most frequent ones. The NER and POS tags are obtained from the Spacy [2] library Tagger model for Italian. The tagger model is run on the sentence word sequence and returns the corresponding NER/POS tags sequence. The custom

[1] The model is implemented in Python 3.6 based on the Keras (keras-gpu 2.1.6 - `https://keras.io/`) library with the Tensorflow open-source deep learning framework (tensorflow 1.8.0 - `https://www.tensorflow.org/`) as backend.

[2] Spacy 2.0.11 - `https://spacy.io/`

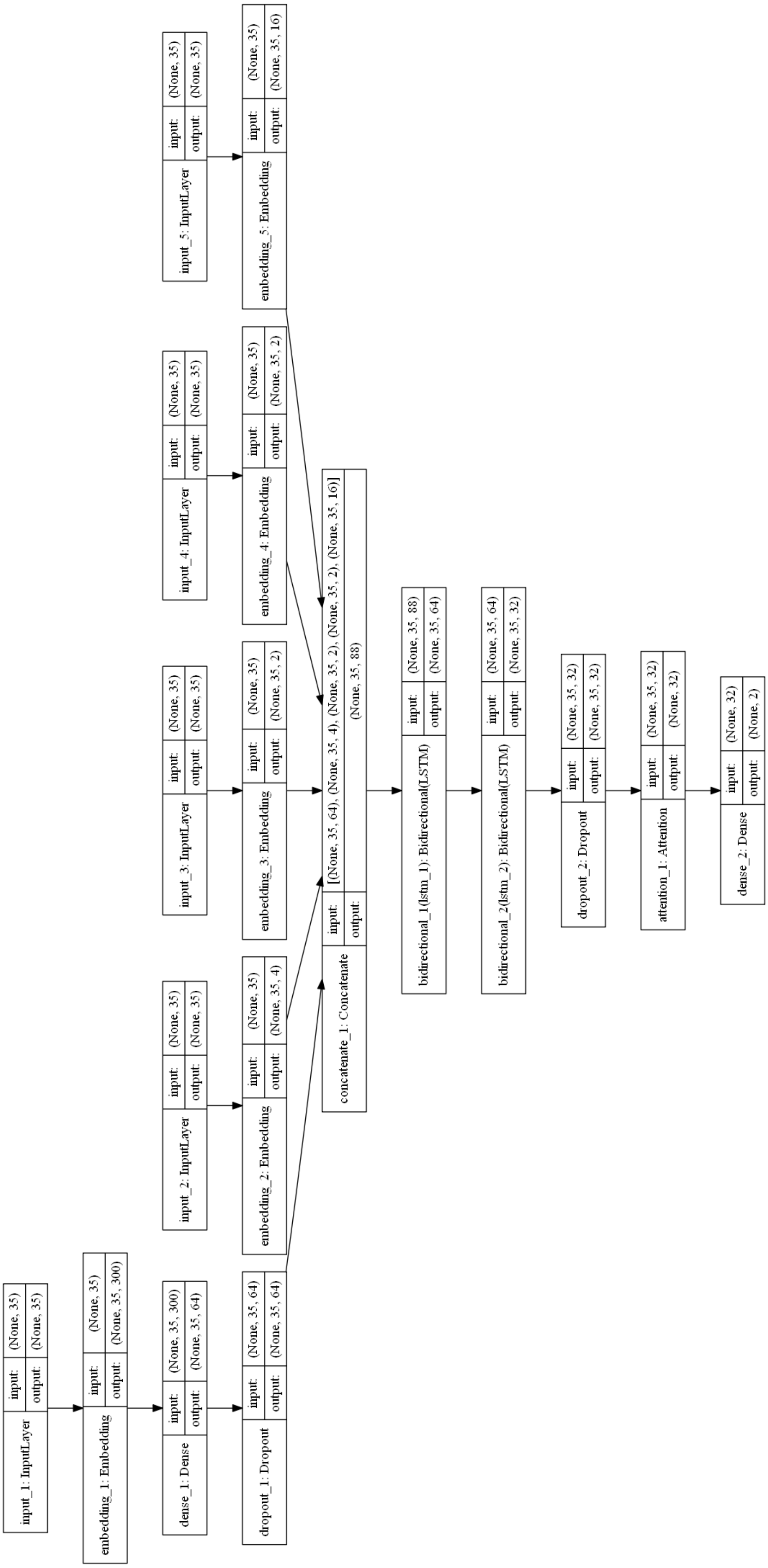


FIGURE 8.1: SentITA model architecture

low dimensional word embeddings are generated by random initialization and are included to provide an embedding representation of the words that are missing from the Fastext embeddings, which otherwise would all be represented by the same Out Of Vocabulary token (OOV token). Moreover, it could be possible to train and fine-tune these custom embeddings on specific datasets to let the model learn the words usage in specific contexts. The information extracted from the OpeNER Sentiment Lexicon Italian are the word polarity with its confidence and they are concatenated in a vector of length 2 that is one of the input to the first layer of the network. The NER tags and POS tags instead are mapped to randomly initialized embeddings of dimensionality respectively 2 and 4 that are not trained during the model training for the ABSITA 2018 task submission. With more data available probably it would be beneficial to train all the NER, POS and custom embeddings but for this specific dataset the results were comparable and slightly better when not training the embeddings.

The model, whose architecture is schematized in Figure 8.1, performs in its initial layer a dimensionality reduction of the Fastext embeddings and then concatenates them with the rest of the embeddings (polarity, NER tag, POS tag, and custom word embeddings) for each each timestep (word) of the sentence sequence. The tensor resulting from the concatenation of the embeddings is fed in a sequence of two bidirectional recurrent layers with LSTM cells. The result of these recurrent layers is passed to the attention mechanism presented in [Raffel et al. 2016]. The attention mechanism in this formulation, produces a fixed-length embedding of the input sequence by computing an adaptive weighted average of the sequence of states (normally denoted as "h") of the RNN. This form of integration is similar to the "global temporal pooling" described in [Sander 2014], which is based on the "global average pooling" technique of [Min et al. 2014]. Finally the output of the attention mechanism goes to the dense output layer (or layers) of the network. The output structure of the network varies depending on the task to which the model is applied. For the ABSITA 2018 challenge they are the aspect detection and aspect polarity signals, while the model included in the SentITA package provides only a sentence sentiment polarity signal. The non linear activations used in the model are Rectified Linear Units (ReLU) for the internal dense layers, hyperbolic tangent (tanh) in the recurrent layers and sigmoid activations in the output dense layer. In order to contrast overfitting the dropout mechanism has been used after the Fastext embedding dimensionality reduction with rate 0.5, in both the recurrent layers between each sequence timestep with rate 0.5 and on the output of the recurrent layers with rate 0.3.

The model has 61,368 trainable parameters and a total of 45,233,366 parameters, the majority of them representing the Fastext embedding matrix (45,000,300). Compared to many NLP models used today the number of trainable parameters is quite small to reduce the possibility of overfitting the training dataset and also because is compensated by the addition of engineered features like polarity dictionary, NER tag and POS tag that help in classifying the examples.

8.4 Data

The data available to train the model are given by a combination of datasets for Italian sentiment analysis. They come from two sources SENTIPOLC 2016 (SENTIment POLarity Classification) and ABSITA 2018 (Aspect-based Sentiment Analysis at EVALITA). They are both subtask of EVALITA, a periodic evaluation campaign of Natural Language Processing and speech tools for the Italian language. The aim of EVALITA is to provide a shared framework where different systems and approaches can be evaluated in a consistent manner.

The main goal of SENTIPOLC 2016 is sentiment classification at message level on Italian tweets. The data we are interested in here, come from the polarity task, that requires, given a message, to predict whether the message is positive, negative, neutral or contains mixed sentiment (i.e. conveying both a positive and negative sentiment). The SENTIPOLC 2016 training dataset contains 7,396 labelled examples.

ABSITA 2018 is an evolution of SENTIPOLC 2016 that aims at capturing the aspect-level opinions expressed in natural language texts in Italian reviews coming from the "Booking.com" website. In this challenge, given the review text, the goal is to identify the "aspect categories" evoked in a sentence and to assign polarity labels to each of the aspect category. The original ABSITA 2018 training set consists of 6,338 hand-labelled sentences while the test set consists of 2,718 sentences. The challenge comprises two closely connected subtask: Aspect Category Detection (ACD) and Aspect Category Polarity (ACP).

In the ACD task one or more "aspect categories" evoked in a review sentence are identified (e.g. the "cleanliness" and "staff" categories). In the Aspect Category Polarity (ACP) task, the polarity of each expressed category is recognized (e.g. a positive category polarity could be expressed concerning the "cleanliness" category while it could be negative if considering the staff category).

In the evaluation framework, the set of aspect categories is known and given to the participants, so the ACD task can be seen as a multi-class, non-exclusive classification task where each input text has to be classified as evoking or not each aspect category. The participating systems have to return a binary vector where each dimension corresponds to an aspect category and the values 0 (false) and 1 (true) indicate whether each aspect has been detected or not in the text.

For the ACP task, the input is the review text paired with the set of aspects identified in the text by the ACD subtask, and the goal is to assign polarity labels to each of the aspect category. Two binary polarity labels are expected for each aspect: POS an NEG, indicating a positive and negative sentiment expressed towards a specific aspect, respectively. The two labels are not mutually exclusive: in addition to the annotation of positive aspects (POS:true, NEG:false) and negative aspects (POS:false, NEG:true), there can be aspects with no polarity, or neutral polarity (POS:false, NEG:false). Finally, the polarity of an aspect can also be mixed (POS:true, NEG:true) in cases where both sentiments are expressed towards a certain aspect in a text.

When the system has participated in the ABSITA 2018 challenge the model has been trained only with the dataset made available from the task organizers [Basile et al. 2018]. The model performance related to the ABSITA 2018 task, thus are representative of a training over 6,338 sentences [Nicola 2018]. In this case no further processing of the dataset is necessary.

For developing the SentITA python package instead the system has been trained on both the SENTIPOLC 2016 and ABSITA 2018 data to leverage a higher number of labelled examples. In order to combine the two dataset together it is necessary to align them accordingly. In fact, while the ABSITA task is similar to SENTIPOLC, its dataset structure is different because it has a polarity label for each aspect category. Since we are interested in detecting whether the sentence expresses positive or negative polarity in general towards any kind of entity, we can neglect the aspect related information of the dataset. In this case we just want to assign a positive/negative polarity label or both to sentences regardless of what aspect class was the subject. For this, we reformulate the dataset label assigning a positive polarity to the sentence if any of evoked aspects polarities is positive and a negative one if any of the evoked aspect polarities is negative. If two different aspects are mentioned one with positive polarity and one with negative polarity we assign both positive and negative labels to the sentence. With this modification the ABSITA dataset structure can be aligned with the SENTIPOLC dataset. Combining the ABSITA train set, with the SENTIPOLC train and test set we obtain a dataset with 13,747 labelled examples. In addition, 50 hand labelled sentences have been added in order to provide more examples with negations and particular Italian idiomatic expression.

8.5 Results

The only preprocessing applied to the text is the conversion of each character to its lower case form. Then, the vocabulary of the model is limited to the first 150,000 words of the Fastext embeddings trough a cap on the max number of embeddings, due to memory constraints of the GPU used for training the model. The Fastext embeddings are sorted by descending frequency of appearance in their training corpus, thus the terms contained in the vocabulary coincide approximately with the 150,000 most frequent Italian words. The other words that are left out from this selection are represented in the model high dimensional embeddings (Fastext embeddings) by an out of vocabulary token. However, all the training set words are anyhow included in the custom low dimensional word embeddings; this is done since both our training text and general users text could be quite different from the one on which Fastext embeddings are trained (specially when working with reviews, tweets and social network platforms). In addition the NER-tagging and POS-tagging models for Italian included in the Spacy library are applied to the text to compute the additional NER-tags and POS-tags features for each word of the sentence sequences.

Like for the datasets, there are some slight differences in the model training for participating into the ABSITA 2018 challenge compared to the model training of the SentITA python package. In the two following subsections we expose the two different training setups.

8.5.1 Training for ABSITA 2018 Challenge

To train the model and generate the challenge submission a k-fold cross validation strategy has been applied. The dataset has been divided in 5 folds and 5 different instantiations of the same model (with the same architecture) have been trained selecting each time a different fold as validation set (20%) and the remaining 4 folds as training set (80%). The number of training epochs is defined with the early stopping technique with patience parameter equal to 7. Once the training epochs are completed, the model snapshot that achieved the best validation loss is loaded. At the end of the training phase, the 5 different models have been applied in inference on the test set and their predictions have been averaged together and thresholded at 0.5. The training of five different instantiations of the same model and the averaging of their predictions overcomes the fact that in each k^{th}-fold the model selection based on the best validation loss is biased on the validation fold itself.

Each of the five models is trained minimizing the crossentropy loss on the different classes with the Nesterov Adam (Nadam) optimizer [Dozat 2016] with default parameters ($\lambda = 0.002$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, schedule_decay = 0.004). The Nesterov Adam optimizer is similar to the Adam optimizer [Kingma and Ba 2014] but the momentum is replaced by the Nesterov momentum [Nesterov 1983]. The Adam optimizer combines two algorithms known to work well for different reasons: momentum, which points the model in a better direction in parameter optimization space, and RMSProp, which adapts how far the model goes in that direction on a per-parameter basis. However, Nesterov momentum which can be viewed as a simple modification of the former, increases stability, and can sometimes provide a distinct improvement in performance, superior to momentum [Sutskever et al. 2013].

This system took part to the ABSITA 2018 Challenge under the name "UNIPV" and obtained the 5^{th} place in the ACD task and the 2^{nd} place in the ACP task as reported respectively in Table 8.1 and 8.2. In these tables the performances of the systems participating to the challenge have been ranked by F1-score from the task organizers. In particular, it is interesting the second place in the ACP since the model is more oriented towards polarity classification, for which it has specific dictionaries, more than aspect detection. This is confirmed also from the high precision score obtained from the model in the ACP task, the 2^{nd} highest among the participating systems.

Ranking	**Micro-Precision**	**Micro-Recall**	**Micro-F1-score**
1	0.8397	0.7837	0.8108
2	0.8713	0.7504	0.8063
3	0.8697	0.7481	0.8043
4	0.8626	0.7519	0.8035
5	0.8819	0.7378	0.8035
6	0.898	0.6937	0.7827
7	0.8658	0.697	0.7723
8	0.7902	0.7181	0.7524
9	0.6232	0.6093	0.6162
10	0.6164	0.6134	0.6149
11	0.5443	0.5418	0.5431
12	0.6213	0.433	0.5104
baseline	0.4111	0.2866	0.3377

TABLE 8.1: Task ACD (Aspect Category Detection) ranking. This system score is reported between dashed lines

Ranking	**Micro-Precision**	**Micro-Recall**	**Micro-F1-score**
1	0.8264	0.7161	0.7673
2	0.8612	0.6562	0.7449
3	0.7472	0.7186	0.7326
4	0.7387	0.7206	0.7295
5	0.8735	0.5649	0.6861
6	0.6869	0.5409	0.6052
7	0.4123	0.3125	0.3555
8	0.5452	0.2511	0.3439
baseline	0.2451	0.1681	0.1994

TABLE 8.2: Task ACP (Aspect Category Polarity) ranking. This system score is reported between dashed lines

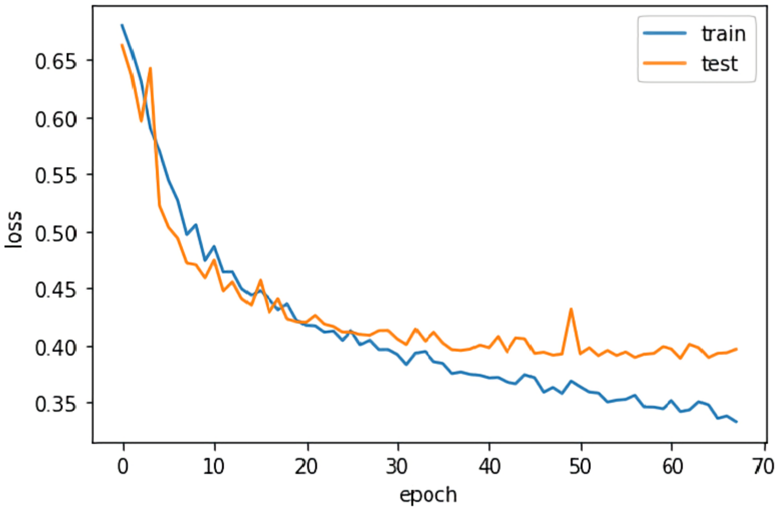


FIGURE 8.2: Model Loss (Categorical Cross Entropy) evolution on train (blue) and test (orange) set across the training epochs

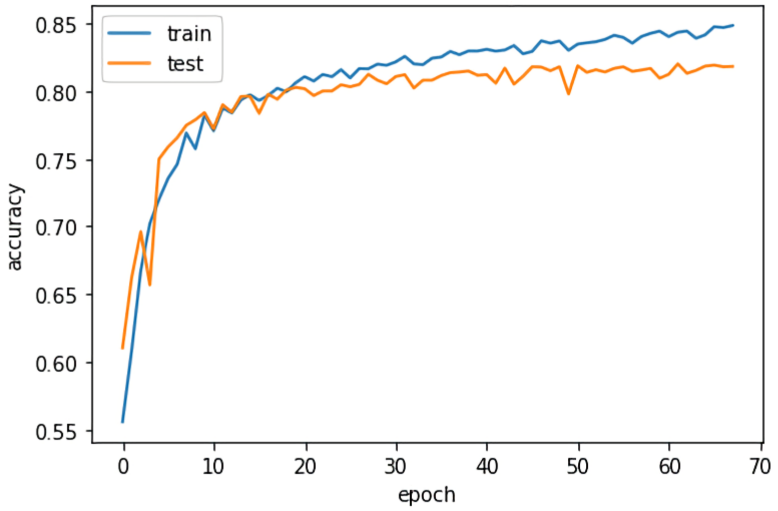


FIGURE 8.3: Model accuracy evolution on train (blue) and test (orange) set across the training epochs

8.5.2 Training of SentITA

When training the model for the SentITA python package implementation, bot the ABSITA and SENTIPOLC datasets are used. For the training of the model, the dataset is split in a train set (85% - 11,727 examples) and a test set (15% - 2,070 examples) after a random shuffling. Then, the model is trained for 67 epochs using early stopping (with $patience = 6$) to reduce overfitting. Finally, the model with the lowest (best) test loss is chosen. In Figure 8.2 and Figure 8.3 is possible to see the evolution of the loss and the accuracy during the training. At the end of the training the lower test loss is 0.389 corresponding to a test accuracy of 82%. We can see that the early stopping is necessary since it halts the training when the two losses start to diverge considerably and the test loss reaches a plateau.

In Table 8.3 we report for inspection some example sentences with the polarity scores computed from the model. As we can see, the model correctly associates the highest scores to the correct polarity. It also correctly handles negations and adverbs like "molto" (very), "grande" (great), "poco" (not much) in connection with adjectives. When the polarity of the

N.	Sentence	Positive	Negative
1	il divano era molto comodo	0.999	0.004
2	il divano non era per niente comodo	0.046	0.982
3	il letto era molto comodo	0.998	0.004
4	il cibo era davvero superbo	0.820	0.454
5	il cibo era davvero buono	0.990	0.024
6	la pasta è cattiva	0.008	0.994
7	il posto era davvero accogliente, i camerieri simpatici e il servizio ottimo, consigliato!	0.999	0.008
8	non è un buon ristorante	0.195	0.923
9	non è stata una cena gradevole	0.191	0.925
10	non è stata una cena per niente gradevole	0.044	0.978
11	sono abbastanza soddisfatto delle prestazioni ma non della batteria	0.767	0.452
12	sono abbastanza soddisfatto delle prestazioni ma non della batteria, dura veramente poco	0.700	0.566
13	sono abbastanza soddisfatto delle prestazioni ma non della batteria, dura veramente poco e scalda un sacco	0.570	0.784
14	è la migliore pasta che abbia mai mangiato	0.806	0.375
15	disastro per i bancari a piazza affari, lasciano 5 punti percentuali sul terreno	0.041	0.499
16	performance positiva dei bancari che centrano il rimbalzo	0.964	0.034
17	domani buone occasioni per gli acquisti grazie ai ribassi dei prezzi per i saldi	0.911	0.024
18	è una canzone con grande musicalità	0.996	0.007
19	è una canzone con poca musicalità	0.173	0.930
20	è un libro avvincente	0.961	0.049
21	è un libro che si fa fatica a leggere	0.081	0.821

TABLE 8.3: SentITA sentiment polarity scores examples

sentence is mixed, the model raises two different signals which in general are also correct in their respective magnitude. The performance can still be improved, for example on polysemy. In fact, it has some difficulty on sentence 4 which is correctly recognized as positive but ideally the positive score should be higher than the one in sentence 5 and the negative one lower. This is due to polysemic words like "superbo" (it can mean superb or also arrogant) that referred to food, it's very positive while referred to a person is considered negative. Also, the model is not perfectly accurate on specific domains like the case of sentence 15 where, while the negative score is definitely larger than the positive one, it is not as high as the positive score in sentence 16. Overall the results of the model are very promising and interesting considering the amount of data used for training and the possibility of further expanding the training set.

8.5.3 SentITA installation and usage

The SentITA package is still in development and there is currently a version made available to the public for local installation. The download is available via Google Drive at the following link[3] and weights approximately 350 Mb. The installation and usage instructions reported in

[3] https://drive.google.com/file/d/1s1BW3T_BysAhVZPai-3AUXpb68aYjQTS/view?usp=sharing

this brief guide are included also within the package archive in the readme file.

How to install SentITA

1. Unzip the downloaded archive
2. cd into the unzipped folder from the console
3. type "pip install ." in the console to install the package locally

How to use SentITA to estimate the polarities of a list of sentences

1. Import the function to calculate the polarity scores with the following code:

   ```
   from sentita import calculate_polarity
   ```

2. define your sentences as a list. e.g.:

   ```
   sentences = ["il viaggio è stato molto interessante",
   "E' la barca a vela più bella che abbia visto",
   "La casa è molto spaziosa e accogliente"]
   ```

3. estimate the sentence polarity by running:

   ```
   results, polarities = calculate_polarity(sentences)
   ```

   "results" is a list of text with the sentence, the positive polarity score and the negative polarity scores. "polarities" is a list of list with the positive and negative polarity score for each sentence, e.g.:
   "polarities[0][0]" contains the positive polarity score of the 1st sentence
   "polarities[2][0]" contains the positive polarity score of the 3rd sentence
   "polarities[2][1]" contains the negative polarity score of the 3rd sentence

8.6 Conclusions

In this chapter the SentITA tool for sentiment analysis in Italian has been presented. The one that has been described is the first iteration of the SentITA python package for general sentiment analysis in Italian. The proposed Bidirectional Attentional LSTM model has been trained on 13,797 examples taken from two publicly available Italian sentiment polarity datasets. The system makes use of different input features that is easy to obtain also through other models like POS and NER tags, polarity embeddings and word embeddings. For this reason, the human effort in the data preprocessing is very limited. The system consistently handles grammar constructions like negations and magnifiers. On the test dataset the model achieves 82% accuracy on the polarity prediction task. Moreover, the results obtained on the ABSITA 2018 challenge are promising, as the system placed 2^{nd} in the ACP and 5^{th} in the ACD task and not very far from the 1^{st} in terms of F1-score.

The model in general shows a high precision but in general a lower recall compared to the other systems. Considering these aspects, the next steps to improve the model performances are mainly in two directions: i) providing more labelled examples to the model and ii) exploiting unsupervised learning. The first can be achieved either hand labelling examples, discovering other available datasets, translating foreign datasets to Italian or identifying sources of text with limited polarity like Wikipedia articles for additional neutral examples.

The second improvement direction would consist in integrating in the model features based on language models or encoder decoder networks. Both additional labels and unsupervised learning would improve the model generalization due to the larger quantity of text available during the training phase. Indeed, covering more topics and lexical content of the Italian language would improve the model recall.

Finally SentITA has been made available through a freely downloadable python package along with a brief guide on its application with the aim of easing future researches that would leverage sentiment analysis in Italian.

Chapter 9

Conclusions

This thesis work investigated the combined use of structured and unstructured (textual) data for systemic risk and bank supervision. Several problems related to these domains have been tackled with different methodologies. In the last chapter, it has also been developed a sentiment polarity classification tool for Italian to ease the analysis of Italian texts in future researches. The investigated models belong to three families: Graphical Gaussian Models, Topic Models and Deep Learning models. All these models have proven to be a valid choice for leveraging numerical (structured) and textual (unstructured) data. Each of them has been applied for solving different problems with different approaches. Graphical Gaussian Models and Topic models have been adopted for inspection and descriptive tasks while deep learning has been applied more for predictive (classification) problems. Throughout the different works presented, the integration of textual (unstructured) and numerical (structured) information has proven useful for systemic risk and bank supervision related analysis. Depending on the task, the integration of textual data has brought either to higher predictive performances or enhanced capability of explaining phenomena and correlating them to other events. In fact, both systemic risk and bank supervision are heavily influenced by the opinions and beliefs that the public and the financial operators form by reading the news.

The valuable information contained in news and other text sources can be challenging to exploit due to dataset specific characteristics, like varying frequency (e.g. tweets) and reliability, and to the intrinsic difficulties in processing natural language. To tackle these difficulties, different strategies have been explored to combine textual and financial data depending on the specific problem. From this cross-section of methodologies and datasets presented in the analysis we can distil some conclusions.

Graphical Gaussian models have proven effective in investigating networks of agents focusing on their connections and mutual correlations. The systemic view that they offer is very useful for systemic risk analysis. In fact, it allows to quantify and consider network effects without requiring too many assumptions on the network structure which often is unknown.

Topic models are very useful for inspecting large text corpora and tracking thematics across time and space dimensions. They are a key tool to understand the composition and characteristics of documents, especially when the corpora are so large that manual inspection is not feasible. In fact, they allow to quickly retrieve the main discussion topics and how they are distributed across the documents. They also allow to group similar documents in clusters and hierarchically organize a collection of texts. Moreover, structural topic models allow to directly take into account additional categorical and continuous variable in the topic recovery process further expanding the possibilities of slicing and dicing the data. For example, adding a time, space or company dimension allows to follow how the discussion topics evolve over time or across different countries and companies.

Deep learning models have given very good results in natural language processing and for classification over high dimensional input space. Their expressive power coupled with the multitude of different architectures available, allows to cover many types of problems. In this work they have been applied for document vector representation, sentiment analysis and classification. When the data availability allows their use, they are a sound choice for handling textual data. In fact, their good performances on high dimensional input spaces like text where each word of the vocabulary can be considered as a variable and they highly non-linear nature allows to model the complexity of natural language.

It's interesting to discuss also the information content of the analyzed types of datasets regarding systemic risk and bank supervision. While each problem is characterized by different data needs, we can draw some useful conclusion on the broad information content of the different data sources and the technical challenges involved in their exploitation.

We examined stocks, macroeconomics and balance sheets data among the structured ones while news articles and micro-blogging texts (Twitter) among the unstructured ones. The research evidences that these data sources hold useful information for systemic risk and bank supervision even if used separately and that their combination has shown to improve performances and problem understanding.

Stock data resulted very helpful for analysis that primarily ground on raw market sentiment towards financial institutions. This has been the case both when investigating correlations and network effects among the different institutions and when exploring the relation among market and crowd sentiment. They can be regarded as one of the main building blocks of systemic risk and bank supervision related analysis given the markets' efficiency and effectiveness in representing companies market values. They allow to perform analysis with an ample timespan dating back far in time for many institutions.

Macroeconomic and balance sheets data allow to include a structural point of view into the analysis. They have proven useful by carrying statistics on the financial institutions' fundamentals and on the environment in which they operate (e.g. country, economic conjuncture). Their information content complements both market and news data offering a different perspective of the same actors. Many aspects of bank supervision depend on and reflect themselves directly on banks balance sheets structures and indirectly on macroeconomic conditions. From a technical point of view when integrating them with other data types there is to consider the lower frequency (e.g. compared to stock data) and the possible discrepancies among different time periods, geographies and institutions. It is especially true for balance sheet data where differences in accounting regulations or practices can create misalignments in the data. Despite all these technical complications to overcome, the combination of these information with other types of data like news has shown very promising results.

News data in general hold an information content that it's complementary to financial structured datasets. As the word says, they regard new events and recent changes that, especially in economics, can act as market drivers. This is true also for systemic risk and bank supervision, where rumours and news regarding financial institutions and the economy impact the financial system stability. They are very interesting for their timeliness and frequency which virtually allows to capture information as soon as it's generated for the public or while it's still propagating. Not only, it's also appealing from a supervisory point of view to track the thematics around which the financial discussions gravitate. This in fact, allows to reconstruct a picture of the relevant topics across time, geographies or other variables that can evidence elements of contagion and network effects.

Twitter data share many characteristics with news data and they could be considered in some sense a subset of them. From a technical point of view, tweets provide both advantages and disadvantages compared to classic news. On one side they consist of shorter and simpler texts focusing only on one argument (for the majority). On the other side the twitter jargon is more difficult to interpret due to slang, abbreviations and implied context. A fascinating aspect of twitter is that it captures a multitude of opinions from every user that interacts on a certain discussion topic. While this plurality can introduce some noise, it allows also to weight the different opinions and sentiment of the crowd towards an argument.

Regarding contagion and systemic risk this can be particularly useful to gauge the sentiment of the crowd and of the economic operators towards financial institutions and themes that can impact financial stability.

Another aspect to be taken in consideration when analysing twitter data, and other textual data sources, is the language plurality. Both tweets and news regarding a topic can be of mixed languages, each one of with its own specificities and required models. In the economic domain, thanks to its internationality, this aspect is mitigated and most of the relevant events are covered also in English by the news providers. Anyway, local and national related events often are better covered and more extensively discussed in the local language. Considering this, an additional barrier to the interpretation of textual data is the dataset and resources availability for the different languages. Widely spoken languages like English, Spanish and Chinese have an advantage over the others in these terms. In fact, in addition to the fact that the interpretation of natural language from machines is an open research field, languages with fewer and smaller labelled dataset available are disadvantaged due to the higher difficulty in training models. For this reason, since many of the works presented in this manuscript make use of sentiment analysis, in the last chapter it has been developed a sentiment classification tool for Italian. Sentiment has been used to interpret and add structure to the textual data (in terms of positive and negative sentiment towards a subject) thus, it is of relevant importance having a model that reliably performs this classification. For this task the use of deep learning models based on word vectors representation has shown very good results. With the application of deep learning in the last years there have been great developments in the natural language processing field, still the interpretation of unstructured text is an unsolved problem. Looking ahead an important role will be played by unsupervised learning models that leverage unlabelled datasets. These models in fact, help reducing the gap between languages with many labelled datasets available and those without and hold promising results for the future. This aspect it's even more relevant for studies that combine textual data with other data types in specific domains, like finance. In these cases, in fact, it's even more rare to have access to large labelled textual dataset on the specific domain.

The investigation of the aforementioned models in combination with the considered data sources has allowed to develop different useful methodologies within the domains of systemic risk and bank supervision. The combination of market and tweet data in graphical models in Chapter 3 has enabled to develop a systemic risk estimation model that has been applied to the Italian banking system. The use of a fast inference algorithm for graphical models in Chapter 4 has brought to a framework for relating information theory measures derived from graphical models to financial stress indexes. The application of Structural Topic Models to a dataset of financial news in Chapter 5 enabled to the track the evolution of thematics over time and follow their spreading among countries. The Granger Causality analysis performed in Chapter 6 between banks market data and tweets sentiment data has shed light on the mutual influence between the two where still market data seem to be prevalent. The combination of news data with bank, sector and macroeconomic level data in Chapter 7 allowed to improve bank distress prediction performance. Finally, the sentiment analysis model develop in Chapter 8 could be a tool to ease similar researches where Italian sentiment analysis is used.

To conclude, we had many positive evidences on the benefits of integrating different types of data, in particular from the inclusion of textual data sources. The complexity of today economic interactions is such that the phenomena are better explained when considered from multiple points of view. Selecting complementary data sources and integrating them in the analysis allows to benefit from this plurality. We believe that the integration of different data types will be an important area of research and applications for the economic domain (also in the general field of machine learning research) in the years to come. The increasing availability of data and computational methods will allow to better exploit the complementary information contained in multiple data types. Considering our specific case, many improvements can still be achieved in the interpretation of text. In this regard, the progresses achieved during the last years in the NLP field and the speed at which new ideas are developed are extremely encouraging to further pursue and expand this research direction.

Bibliography

[Acharya et al. 2010] Acharya, V., Pedersen, L., Philippon, T. and Richardson, M. (2010). Measuring systemic risk. Working paper, Federal Reserve of Cleveland.

[Acharya et al. 2012] Acharya, V., Engle, R., and Richardson M. (2012). Capital shortfall: a new approach to ranking and regulating systemic risks. American Economic Review, 102(3), pages 59-64.

[Adrian and Brunnermeier 2009] Adrian, T. and Brunnermeier M. (2009). CoVaR. Technical report, Princeton University.

[Ahelegbey et al. 2015] Ahelegbey, D., Billio, M. and Casarin, R. (2015). Bayesian Graphical Models for Structural Vector Autoregressive Processes. Journal of Applied Econometrics.

[Antenucci et al. 2014] Antenucci, D., Cafarella, M., Levenstein, M. C., R, C., and Shapiro, M. (2014). Using social media to measure labor market flows.

[Attardi et al. 2016] Attardi, G., Sartiano, D., Alzetta, C. and Semplici, F. (2016). Convolutional Neural Networks for Sentiment Analysis on Italian Tweets. CLiC-it/EVALITA.

[Bakker 2002] Bakker, B. (2002). Reinforcement Learning with Long Short-Term Memory. In Advances in Neural Information Processing Systems, 14.

[Banulescu and Dumitrescu 2015] Banulescu, G. and Dumitrescu, E. (2015). Which are the SIFIs? A component expected shortfall approach to systemic risk. Journal of Banking and Finance, Volume 50 (2015): 575-588.

[Barber and Odean 2008] Barber M. and Odean, T. (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. The Review of Financial Studies, 21(2):785–818.

[Barbieri et al. 2016] Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N. and Patti, V. (2016). Overview of the Evalita 2016 SENTIment POLarity Classification Task. In Proceedings of CLiC-it 2016 and EVALITA 2016, Napoli, Italy, December 5-7, 2016. Volume 1749 of CEUR Workshop Proceedings. CEUR-WS.org.

[Barfuss et al. 2016] Barfuss, W., Massara, G., Di Matteo, T. and Aste, T., (2016). "Parsimonious modeling with information filtering networks". Physical Review E, 94 (6). ISSN 1539-3755

[Barigozzi et al. 2013] Barigozzi, M. and Brownlees, C. (2013). Nets: Network Estimation for Time Series, No. 723, Working Papers, Barcelona Graduate School of Economics.

[Barrett et al. 2010] Barrett, A., Barnett, L. and Seth, A. (2010). Multivariate Granger causality and generalized variance. Physical Review E 81, pages 041907.

[Basile and Nissim 2013] Basile V. and Nissim, M. (2013). Sentiment Analysis on Italian Tweets. 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 100–107.

[Basile et al. 2018] Basile, P., Basile, V., Croce, D. and Polignano, M. (2018). Overview of the EVALITA Aspect-based Sentiment Analysis (ABSITA) Task. Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)

[Battiston et al. 2012] Battiston S., Puliga M., Kaushik R., Tasca P. and Caldarelli G. (2012). 'DebtRank: Too Central to Fail? Financial Networks, the FED and Systemic Risk', Scientific Reports, 2, 541.

[Bekaert et al. 2005] Bekaert G., Harvey, C. and Ng., A. (2005). Market integration and contagion. Journal of Business, 78(1): 39-70.

[Bengio et al. 1994] Bengio, Y., Simard, P. and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2):157–166

[Bengio et al. 2003] Bengio, Y., Ducharme, R., Vincent, P. and Janvin, C. (2003). A neural probabilistic language model. The Journal of Machine Learning Research, 3:1137–1155.

[Benoit et al. 2015] Benoit, S., Colliard, J., Hurlin, C. and Perignon, C. (2015). Where the Risks Lie: A Survey on Systemic Risk. HEC Paris Research Paper No. FIN–2015–1088.

[Berger 1985] Berger, J. (1985). Statistical Decision Theory and Bayesian Analysis, New York: Springer-Verlag.

[Betz et al. 2003] Betz, F., Opricǎ, S., Peltonen, T. and Sarlin, P. (2014). Predicting distress in European banks. Journal of Banking & Finance, 45:225–241.

[Bholat et al. 2015] Bholat, D., Hansen, S., Santos, P. and Schonhardt-Bailey, C. (2015). Text mining for central banks. In Centre for Central Banking Studies Handbook, volume 33. Bank of England.

[Billio et al. 2012] Billio, M., Getmansky, M., Lo, A., Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sector. Journal of Financial Economics, 104(3), 535-559.

[Bing 2012] Bing, L. (2012). Sentiment analysis: A fascinating problem. In Sentiment Analysis and Opinion Mining, pages 7–143. Morgan and Claypool Publishers.

[Bishop 2006] Bishop, C. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.

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

[Blei and Lafferty 2006] Blei, D. and Lafferty, J. (2006). Correlated topic models. In Y. Weiss, B. Scholkopf, and J. Platt, editors, Advances in Neural Information Processing Systems 18, pages 147–154. MIT Press, Cambridge, MA.

[Blei and Lafferty 2009] Blei, D. and Lafferty, J. (2009). "Topic Models." In A Srivastava, M Sahami (eds.), Text Mining: Classification, Clustering, and Applications. Chapman & Hall/CRC Press.

[Blei et al. 2003] Blei, D., Ng, A. and Jordan, M. (2003). Latent Dirichlet allocation. The Journal of Machine Learning Research, 3:993–1022. ISSN 1532-4435.

[Bojanowski et al. 2016] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2016) Enriching Word Vectors with Subword Information. arXiv:1607.04606v2.

[Bollen et al. 2011] Bollen, J., Mao, H. and Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1):1–8.

[Bordino et al. 2012] Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A. and Weber, I. (2012). Web search queries can predict stock market volumes. PloS one, 7(7), e40014.

[Bosco et al. 2013] Bosco C., Patti, V. and Bolioli, A. (2013). Developing corpora for sentiment analysis: The case of irony and senti-tut. IEEE Intelligent Systems, 28(2).

[Boss et al. 2004] Boss, M., Elsinger, H., Summer, M. and Thurner, S. (2004). Network topology of the interbank market, Quantitative Finance, 4:6, 677-684.

[Bottou 1998] Bottou, L. (1998). Online Algorithms and Stochastic Approximations, Online Learning and Neural Networks, Edited by David Saad, Cambridge University Press, Cambridge, UK.

[Breiman 1994] Breiman, L. (1994). Bagging predictors. Machine Learning, 24(2), 123–140.

[Brown 2012] Brown, E. (2012). Will Twitter make you a better investor? A look at sentiment, user reputation and their effect on the stock market. In Proceedings of the Southern Association for Information Systems Conference. Atlanta: SAIS, pages 36–42.

[Brown et al. 1992] Brown, P., Della Pietra, V., deSouza, P., Lai, J. and Mercer, L. (1992). Class-based n-gram models of natural language. Computational Linguistics, 18(4):467–479.

[Brownlees and Engle 2011] Brownlees, C. and Engle, R. (2011). Volatility, correlation and tails for systemic risk measurement. Technical report, New York University.

[Brunnermeier and Oehmke 2012] Brunnermeier, M. and Oehmke, M. (2012). Bubbles, Financial Crises, and Systemic Risk. NBER Working Papers 18398, National Bureau of Economic Research.

[Bryson et al. 1963] Bryson, A., Denham, W. and Dreyfus, S. (1963). Optimal programming problems with inequality constraints I: necessary conditions for extremal solutions. AIAA Journal, 1:2544-2550.

[Buntine 1995] Buntine, W. (1995). Chain graphs for learning. In Proceedings of the Conference on Uncertainty in Artificial Intelligence.

[Calabrese and Giudici 2015] Calabrese, R. and Giudici, P. (2015). Estimating bank default with generalised extreme value models. Journal of the Operational Research Society.

[Cao 2013] Cao, Z. (2013). Multi-CoVaR and Shapley Value: A Systemic Risk Measure. Working Paper. Banque de France DSF-SMF.

[Carlin and Louis 2000] Carlin, B., Louis, T. (2000). Bayes and Empirical Bayes Methods for Data Analysis (2nd ed.). Chapman & Hall/CRC.

[Casella 1985] Casella, G. (1985). An Introduction to Empirical Bayes Data Analysis. American Statistician (American Statistical Association) 39(2), 83-87.

[Castellucci et al. 2016] Castellucci, G., Croce, D. and Basili, R. (2016). Context–aware Convolutional Neural Networks for Twitter Sentiment Analysis in Italian. CLiC-it/EVALITA.

[Cerchiello and Giudici 2015] Cerchiello, P. and Giudici, P. (2015). How to measure the quality of financial tweets. Quality & Quantity, 50(4), pages 1-19.

[Cerchiello and Giudici 2016] Cerchiello, P. and Giudici, P. (2016). Conditional graphical models for systemic risk estimation. Expert systems with applications. Vol. 43, pages 165-174.

[Cerchiello and Giudici 2016b] Cerchiello, P. and Giudici, P. (2016). Big data analysis for financial risk management. Journal of Big Data, Vol. 3:18.

[Cerchiello et al. 2017] Cerchiello, P., Giudici, P. and Nicola, G. (2017). Twitter data models for bank risk contagion, In Neurocomputing, Volume 264, Pages 50-56.

[Cerchiello et al. 2017b] Cerchiello, P., Nicola, G., Rönnqvist, S. and Sarlin, P. (2017). Deep Learning Bank Distress from News and Numerical Financial Data. DEM Working paper. arXiv:1706.09627.

[Chawla et al. 2016] Chawla, N., Da, Z., Xu, J. and Ye, M. (2016). Information Diffusion on Social Media: Does It Affect Trading, Return, and Liquidity?. Working paper - University of Notre-dame.

[Chen and Chaudhari 2005] Chen, J. and Chaudhari, N. (2005). Protein Secondary Structure Prediction with bidirectional LSTM networks. In International Joint Conference on Neural Networks: Post-Conference Workshop on Computational Intelligence Approaches for the Analysis of Bio-data (CI-BIO).

[Cho 2015] Cho, K. (2015). Natural language understanding with distributed representation. arXiv preprint arXiv:1511.07916.

[Cho et al. 2014] Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H. and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In EMNLP.

[Choi and Varian 2012] Choi, H. and Varian, H. (2012). Predicting the present with google trends. Economic Record, 88(s1), 2-9.

[Choromanska et al. 2015] Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. (2015). The loss surfaces of multilayer networks. In AISTATS.

[Cimino et al. 2016] Cimino, A. and Dell'Orletta, F. (2016). Tandem LSTM–SVM Approach for Sentiment Analysis. CLiC-it/EVALITA.

[Clark 2003] Clark, A. (2003). Combining distributional and morphological information for part of speech induction. In Proceedings of EACL.

[Collobert and Weston 2008] Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning, pages 160–167.

[Collobert et al. 2011] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research, 12:2493-2537.

[Constantin et al. 2016] Constantin, A., Peltonen, T. and Sarlin, P. (2016). Network linkages to predict bank distress. Journal of Financial Stability.

[Corsetti et al. 2001] Corsetti, G., Pericoli, M., Sbracia, M. (2001). Correlation analysis of financial contagion: what one should know before running a test. Temi di Discussione. No. 408, June, Bank of Italy.

[Cybenko 1989] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Math. Control Signal Systems 2: 303.

[Da et al. 2011] Da, Z., Engelberg, J. and Gao, P. (2011) In search of attention. The Journal of Finance, 66 (5):1461–1499.

[Darroch et al. 1980] Darroch, J., Lauritzen, S. and Speed, T. (1980). Markov fields and log-linear interaction models for contingency tables. The Annals of Statistics, pages 522–539.

[Dauphin et al. 2014] Dauphin, Y., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional nonconvex optimization. In Advances in neural information processing systems, pages 2933–2941.

[Dawid and Lauritzen 1993] Dawid, A., Lauritzen, S. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. Annals of Statistics, 21, 1272-1317.

[De Bandt and Hartmann 2000] De Bandt, O. and Hartmann, P. (2000). Systemic Risk: A Survey. ECB Working Paper No. 35. November 2000.

[Deerwester et al. 1990] Deerwester, S., Dumais, S., Furnas, G., Landauer, T. and Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391–407.

[Deloche 2017] Deloche, F. (2017). A diagram for a one-unit recurrent neural network (RNN). In Wikimedia Commons.

[Dempster et al. 1977] Dempster, A., Laird, N. and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1), pages 1-38.

[Deriu et al. 2016] Deriu, J. and Cieliebak, M. (2016). Sentiment Detection using Convolutional Neural Networks with Multi–Task Training and Distant Supervision. CLiC-it/EVALITA.

[Deutsche Bundesbank 2014] Deutsche Bundesbank (2014). Exchange rates and financial stress. Monthly report, July.

[Diaconis and Ylvisaker 1979] Diaconis, P. and Ylvisaker, D. (1979). Conjugate Priors for Exponential Families. Ann. Statist. 7, no. 2, 269-281.

[Diamond and Dybvig 1983] Diamond, D. and Dybvig, P. (1983). Bank runs, deposit insurance, and liquidity". Journal of Political Economy. 91 (3): 401–419.

[Diebold and Yilmaz 2014] Diebold, F. and Yilmaz, K. (2014). On the Network Topology of Variance Decompositions: Measuring the Connectedness of Financial Firms. Journal of Econometrics, 182, 119-134.

[Dilip et al. 2013] Dilip, K., Patro, M. and Xian, S. (2013). A simple indicator of systemic risk. Journal of Financial Stability, Volume IX, pages 105 - 116

[Ding et al. 2015] Ding, X., Yue, Z., Ting, L. and Junwen, D. (2015). Deep Learning for Event-Driven Stock Prediction. Paper presented at the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015), Buenos Aires, Argentina, July 25-31.

[Dozat 2016] Timothy, D. (2016). Incorporating Nesterov Momentum into Adam. In Proceedings of 4th International Conference on Learning Representations, Workshop Track, 2016.

[Dung et al. 2008] Dung, L., Komeda, T. and Takagi, M. (2008). Reinforcement learning for POMDP using state classification. Applied Artificial Intelligence, 22(7-8):761–779.

[ECB 2009] ECB (2009). ECB Financial Stability Review, December 2009.

[Eck and Schmidhuber 2002] Eck, D. and Schmidhuber, J. (2002). Finding Temporal Structure in Music: Blues Improvisation with LSTM Recurrent Networks. In H. Bourlard, editor, Neural Networks for Signal Processing XII, Proceedings of the 2002 IEEE Workshop, pages 747–756, New York.

[Eisenberg and Noe 2001] Eisenberg, L. and Noe, T. (2001). Systemic Risk in Financial Networks. Management Science, Vol. 47, Issue 2, pages 236-249.

[Eisenstein et al. 2011] Eisenstein, J., Amr, A. and Xing, E. (2011). Sparse Additive Generative Models of Text. In Proceedings of the 28th International Conference on Machine Learning, pages 1041–1048.

[Elman 1990] Elman, J. (1990). Finding structure in time. Cognitive science, 14(2):179–211.

[Engle 2009] Engle, R. (2009). Anticipating correlations. Princeton, NJ: Princeton University Press.

[Federal Reserve 2010] Kliesen, K. and Smith, D. (2010). Measuring Financial Market Stress. St. Louis Fed Economic Synopses, 2010, n.2.

[Federal Reserve Bank of St. Louis 2018] Federal Reserve Bank of St. Louis (2018). St. Louis Fed Financial Stress Index. Retrieved from FRED.

[Feldman 2013] Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82–89.

[Firth 1957] Firth, J. (1957). A synopsis of linguistic theory. Vol. 1952–1959:1–32, pages 1930–1955.

[Forbes and Rigobon 2002] Forbes, K. and Rigobon, R. (2002). No Contagion, Only Interdependence: Measuring Stock Market Comovements. Journal of Finance, 2002, vol. 57, issue 5, pages 2223-2261.

[Gal and Ghahramani 2016] Gal, Y. and Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural networks. In Advances in neural information processing systems, pages 1019–1027.

[Giannini et al. 2013] Giannini, R., Irvine, P. and Shu, T. (2013). The convergence and divergence of investors' opinions around earnings news: Evidence from a social network. Asian Finance Association (AsFA) 2013 Conference.

[Girolami and Kaban 2003] Girolami, M. and Kaban, A. (2003). On an Equivalence between PLSI and LDA. In Proceedings of ACM, pages 433-434.

[Giudici 2001] Giudici, P. (2001). Bayesian data mining, with application to financial benchmarking and credit scoring. Applied stochastic models in business and industry, 17, 69-81.

[Giudici and Green 1999] Giudici, P. and Green, P. (1999). Decomposable graphical Gaussian model determination. Biometrika, 86, 785–801.

[Giudici and Spelta 2016] Giudici, P. and Spelta, A. (2016). Graphical network models for international financial flows (2016). Journal of Business and Economic Statistics.

[Glorot and Bengio 2010] Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deepfeedforward neural networks. In AISTATS.

[Goldberg 2015] Goldberg, Y. (2015). A primer on neural network models for natural language processing. arXiv preprint arXiv:1510.00726.

[Goodfellow et al. 2015] Goodfellow, I., Vinyals, O. and Saxe, A. (2015). Qualitatively characterizing neural network optimization problems. In International Conference on Learning Representations (ICLR 2015).

[Goodfellow et al. 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning. MIT Press. http://www.deeplearningbook.org.

[Gorton 1988] Gorton, G. (1988). Banking Panics and Business Cycles. Oxford Economic Papers, 40, pages 751-781.

[Granger 1969] Granger, C. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica, 37(3):424–438.

[Granger 1988] Granger, C. (1988). Some Recent Developments in a Concept of Causality. Journal of Econometrics, 39:199–211.

[Grave et al. 2018] Grave, E., Bojanowski, P., Gupta, P., Joulin, A. and Mikolov, T. (2018) Learning Word Vectors for 157 Languages. Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).

[Graves and Schmidhuber 2005] Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5):602–610.

[Graves et al. 2006] Graves, A., Fernandez, S., Gomez, F. and Schmidhuber, J. (2006). Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In Proceedings of the International Conference on Machine Learning (ICML 2006).

[Graves et al. 2008] Graves, A., Fernandez, S., Liwicki, M., Bunke, H. and Schmidhuber, J. (2008). Unconstrained Online Handwriting Recognition with Recurrent Neural Networks. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, Advances in Neural Information Processing Systems 20. MIT Press, Cambridge, MA.

[Griffiths and Steyvers 2004] Griffiths, T. and Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl. 1):5228–5235.

[Guo et al. 2008] Guo S., Seth A., Kendrick K., Zhou C., Feng J. (2008). Partial Granger causality - eliminating exogenous inputs and latent variables. Journal of Neuroscience Methods 172(1): 79-93.

[Guyon and Elisseeff 2003] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, vol. 3, pages 1157-1182.

[Hahnloser and Seung 2001] Hahnloser, R. and Seung, H. (2001). Permitted and forbidden sets in symmetric threshold-linear networks. In Advances in Neural Information Processing Systems, pages 217–223.

[Hahnloser et al. 2000] Hahnloser, R., Sarpeshkar, R., Mahowald, M., Douglas, R. and Seung, H. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature, 405(6789):947–951.

[Hammer 2000] Hammer, B. (2000). On the approximation capability of recurrent neural networks. Neurocomputing, 31(1):107–123.

[Harris 1954] Harris, Z. (1954). Distributional structure. Word, 10:146–162

[Hautsch et al. 2014] Hautsch, N., Schaumburg, J. and Schienle, M. (2014) Forecasting Systemic Impact in Financial Networks. International Journal of Forecasting, 30, 781-794.

[He et al. 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision, pages 1026–1034.

[Heredia et al. 2016] Heredia, B., Khoshgoftaar, T., Prusa, J. and Crawford, M. (2016). CrossDomain Sentiment Analysis: An Empirical Investigation, 2016 IEEE 17th International Conference on Information Reuse and Integration, pages 160-165.

[Hinton et al. 1986] Hinton, G., McClelland, J., and Rumelhart, D. (1986). Distributed representations. In Rumelhart, D. E. and McClelland, J. L., editors, Parallel Distributed Processing: Explorations in the Microstructure of Cognition. 1986. Volume 1: Foundations, MIT Press, Cambridge, MA. pages 77-109.

[Hinton et al. 2012] Hinton, G., Srivastava, N., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. CoRR, abs/1207.0580.

[Hochreiter 1991] Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, TU Munich.

[Hochreiter and Schmidhuber 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation 9(8):1735-1780.

[Hochreiter et al. 2007] Hochreiter, S., Heusel, M. and Obermayer, K. (2007). Fast Model-based Protein Homology Detection without Alignment. Bioinformatics.

[Hodrick and Prescott 1980] Hodrick, R. and Prescott, E. (1980). Post-war U.S. business cycles: An empirical investigation. Mimeo. Carnegie-Mellon University, Pittsburgh, PA.

[Hofmann 1999] Hofmann, T. (1999). Probabilistic latent semantic indexing. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99), pages 50–57.

[Hokkanen et al. 2015] Hokkanen, J., Jacobson, T., Skingsley, C. and Tibblin, M. (2015). The Riksbank's future information supply in light of Big Data. In Economic Commentaries, volume 17. Sveriges Riksbank.

[Howard et al. 2018] Howard, J., Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. https://doi.org/arXiv:1801.06146v3.

[Hu M., Liu B. 2004] Hu, M., Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004).

[Huang et al. 2011] Huang, X., Zhou, H., Zhu, H. (2011). Systemic risk contribution. Technical report, Board of Governors of the Federal reserve System.

[Hyndman et al. 2008] Hyndman, R. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. Journal of Statistical Software, 26(3), 1–22.

[Hyndman et al. 2018] Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E. and Yasmeen, F. (2018). forecast: Forecasting functions for time series and linear models. R package version 8.4.

[Idier et al. 2013] Idier, J., Lame', G. and Mesonnier, J. (2013). How useful is the marginal expected shortfall for the measurement of systemic exposure? a practical assessment. Working paper series, 1546, European Central Bank.

[Ivakhnenko 1967] Ivakhnenko, A. and Grigor'evich, L. (1967). Cybernetics and forecasting techniques. American Elsevier Pub. Co., NY.

[Jensen 1996] Jensen, F. (1996). An introduction to Bayesian networks, volume 36. UCL press London.

[Jordan et al. 1999] Jordan, M., Ghahramani, Z., Jaakkola, T. and Saul, L. (1999). An introduction to variational methods for graphical models. Machine learning, 37(2):183–233.

[Kalchbrenner et al. 2014] Kalchbrenner, N., Grefenstette, E. and Blunsom, P. (2014). A Convolutional Neural Network for Modelling Sentences. In Proceedings of ACL.

[Kaminsky et al. 2003] Kaminsky, G., Reinhart, C. and Végh, C. (2003). The Unholy Trinity of Financial Contagion. Journal of Economic Perspectives, 17 (4), pages 51-74.

[Karolyi and Stulz 2006] Karolyi, A. and Stulz, R. (1996). Why Do Markets Move Together? An Investigation of U.S.-Japan Stock Return Comovements. Journal of Finance 51, pages 951–986.

[Kendall and Stuart 1979] Kendall, M. and Stuart, A. (1979). The Advanced Theory of Statistics: Inference and Relationship. Hodder Arnold, London.

[Kindermann et al. 1980] Kindermann, R. and Snell J. (1980). Markov random fields and their applications. American Mathematical Society Providence, RI.

[Kingma and Ba 2014] Kingma, D. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. International Conference on Learning Representations. https://arxiv.org/pdf/1412.6980.pdf.

[kiros et al. 2015] Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R., Torralba, A., Urtasun, R., and Fidler, S. (2015). Skip-thought vectors. Advances in Neural Information Processing Systems (NIPS).

[Koller and Friedman 2009] Koller, D. and Friedman, N. (2009). Probabilistic graphical models: principles and techniques. The MIT Press.

[Landauer and Dumais 1997] Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211-240.

[Landauer et al. 1998] Landauer, T., Foltz, P. and Laham, D. (1998). Introduction to Latent Semantic Analysis. Discourse Processes. 25 (2–3): 259–284, 1998.

[Lauritzen 1996] Lauritzen, S. (1996). Graphical models. Oxford University Press.

[Lauritzen and Wermuth 1989] Lauritzen, S. and Wermuth N. (1989). Graphical models for association between variables, some of which are qualitative and some quantitative. Annals of Statistics, 17(1):31–57.

[Le and Mikolov 2014] Le, Q. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. Proceedings of the 31 st International Conference on Machine Learning, Beijing, China. JMLR: W&CP, volume 32.

[LeCun et al. 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature, 521(7553):436–444.

[Lindgren et al. 1996] Lindgren, C., Garcia, G. and Saal, M. (1996). Bank Soundness and Macroeconomic Policy. (Washington, DC: International Monetary Fund).

[Liu 2015] Liu, B. (2015). Sentiment analysis: mining opinions, sentiments, and emotions. The Cambridge University Press.

[Liwicki et al. 2007] Liwicki, M., Graves, A., Fernandez, S., Bunke, H. and Schmidhuber, J. (2007). A Novel Approach to On-Line Handwriting Recognition Based on Bidirectional Long Short-Term Memory Networks. In Proceedings of the 9th International Conference on Document Analysis and Recognition, ICDAR 2007.

[Mackay 2003] MacKay, D. (2003). Information Theory, Inference, and Learning Algorithms. Cambridge University Press, 2003.

[Malo et al. 2014] Malo, P., Sinha, A., Korhonen, P., Wallenius, J. and Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology, 65(4):782–796.

[Mantegna 1999] Mantegna, R. (1999). Hierarchical structure in financial markets. The European Physical Journal B: Condensed Matter and Complex Systems, vol. 11, issue 1, pages 193–197.

[Mao et al. 2015] Mao, H., Countsi, S. and Bollen, J. (2015). Quantifiying the Effects of online bullishness on International Financial Markets. ECB Statistics Paper Series, 9:1–21.

[Martens and Sutskever 2011] Martens, J. and Sutskever, I. (2011). Training Recurrent Neural Networks with Hessian-Free optimization. ICML.

[Martin et al. 1998] Martin, S., Liermann, J. and Ney, H. (1998). Algorithms for bigram and trigram word clustering. Speech Communication, 24, 19–37.

[Massara et al. 2016] Massara, G., DiMatteo, T. and Aste, T. (2016). Network Filtering for Big Data: Triangulated Maximally Filtered Graph". Journal of Complex Networks, Volume 5, Issue 2, pages 161–178.

[Masson 1998] Masson, P. (1998). Contagion: Monsoonal Effects, Spillovers, and Jumps Between Multiple Equilibria. IMF Working Paper, International Monetary Fund.

[Mayer et al. 2006] Mayer, H., Gomez, F., Wierstra, D., Nagy, I., Knoll, A. and Schmidhuber, J. (2006). A system for robotic heart surgery that learns to tie knots using recurrent neural networks. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 543–548.

[McCulloch and Pitts 1943] McCulloch, W. and Pitts, W. (1943). A Logical Calculus of Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics. 5 (4): 115–133.

[Merton 1974] Merton, R. (1974). On the pricing of corporate debt: the risk structure of interest rates. Journal of Finance, 2, 449–471.

[Mikolov 2012] Mikolov, T. (2012). Statistical Language Models Based on Neural Networks. PhD thesis, PhD Thesis, Brno University of Technology, 2012.

[Mikolov et al. 2013] Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient estimation of word representations in vector space. In Proceedings of Workshop at International Conference on Learning Representations (ICLR 2013).

[Mimno and McCallum 2008] Mimno, D. and McCallum, A. (2008). Topic Models Conditioned on Arbitrary Features with Dirichlet-Multinomial Regression. In Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI2008), Helsinki, Finland, July 9-12.

[Min et al. 2014] Min, L., Qiang, C. and Shuicheng, Y. (2014). Network in network. arXiv preprint arXiv:1312.4400.

[Mittal and Goel 2012] Mittal, A. and Arpit, G. (2012). Stock prediction using twitter sentiment analysis. Standford University, CS229.Stanford.Edu.

[Nair and Hinton 2010] Nair, V. and Hinton, G. (2010). Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), pages 807–814.

[Nakov et al. 2016] Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F. and Stoyanov, V. (2016). SemEval-2016 Task 4: Sentiment Analysis in Twitter. Proceedings of SemEval-2016, pages 1–18.

[Nann et al. 2013] Nann, S., Krauss, J. and Schoder, D. (2013). Predictive Analytics On Public Data—The Case Of Stock Markets. ECIS 2013 Completed Research. 102.

[Nesterov 1983] Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate o (1/k2). In Soviet Mathematics Doklady, volume 27, pages 372-376.

[Nicola 2018] Nicola, G. (2018). Bidirectional Attentional LSTM for Aspect Based Sentiment Analysis on Italian. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18), Turin, Italy. CEUR.org.

[Nielsen 2015] Nielsen, M. (2015). Neural Networks and Deep Learning. Determination Press.

[Nyman et al. 2015] Nyman, R., Gregory, D., Kapadia, K., Ormerod, P., Tuckett, D. and Smith, R. (2015). News and narratives in financial systems: exploiting big data for systemic risk assessment. BoE, mimeo.

[Nymand-Andersen 2016] Nymand-Andersen, P. (2016). Big data: The hunt for timely insights and decision certainty. IFC Working Papers, 14.

[Olah 2015] Olah, C. (2015). Understanding LSTM Networks. In http://colah.github.io.

[Oliveira et al. 2013] Oliveira, N., Cortez, P. and Area, N. (2013). On the predictability of stock market behaviour using stock tweets sentiment and posting volume. In Progress in Artificial Intelligence. EPIA 2013, pages 355-365.

[Onnela et al. 2004] Onnela, J., Kaski, K. and Kertesz, J. (2004). Clustering and information in correlation based financial networks, The European Physical Journal B, 38, pages 353–362.

[Ozgöde 2011] Ozgöde, O. (2011). The Emergence of Systemic Financial Risk: From Structural Adjustment (Back) to Vulnerability Reduction, www.limn.it, Issue number one: Systemic Risk.

[Pang et al. 2002] Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86.

[Pang et al. 2008] Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis.

[Pearl 1999] Pearl, J. (1999). Causality: Models, Reasoning, and Inference. Cambridge University Press.

[Peltonen et al. 2015] Peltonen, T., Piloiu, A. and Sarlin, P. (2015). Network Linkages to Predict Bank Distress.

[Pennignton et al. 2014] Pennington, J., Socher, R. and Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543. ACL.

[Plank et al. 2016] Plank, B., Søgaard, A. and Goldberg, Y. (2016). Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In ACL.

[Putthividhya et al. 2009] Putthividhya, D., Hagai T. and Srikantan, N. (2009). Independent Factor Topic Models. In proceedings of the 26th Annual International Conference on Machine Learning, Montreal, Quebec, Canada, June 14–18; pages 833-840.

[Rönnqvist and Sarlin 2017] Rönnqvist, S. and Sarlin, P. (2017). Bank distress in the news: Describing events through deep learning, Neurocomputing, Volume 264, Pages 57-70.

[Raffel et al. 2016] Raffel, C. and Ellis, D. (2016). Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems. https://arxiv.org/abs/1512.08756.

[Ranco et al. 2015] Ranco, G., Aleksovski, D., Caldarelli, G., Grcar, M. and Mozetic, I. (2015). The Effects of Twitter Sentiment on Stock Price Returns. PLoS ONE 10(9): e0138441.

[Rao and Srivastava 2012] Rao, T. and Srivastava, S. (2012). Twitter sentiment analysis: How to hedge your bets in the stock markets. CoRR, abs/1212.1107.

[Rigobon 2016] Rigobon, R. (2016). Contagion, spillover and interdependence. ECB Working Paper 1975, November 2016.

[Roberts et al. 2016] Roberts, M., Brandon, M. and Tingley, D. (2016). Navigating the Local Modes of Big Data: The Case of Topic Models. In Data Analytics in Social Science, Government, and Industry. New York: Cambridge University Press.

[Roberts et al. 2016b] Roberts, M., Brandon, M. and Airoldi, E. (2016). A model of text for experimentation in the social sciences. Journal of the American Statistical Association 111: 988-1003.

[Roelstraete et al. 2011] Roelstraete, B. and Rosseel, Y. (2011). FIAR: An R Package for Analyzing Functional Integration in the Brain. Journal of Statistical Software, 44(13), 1-32.

[Roelstraete et al. 2012] Roelstraete, B. and Rosseel, Y. (2012) "Does Partial Granger Causality Really Eliminate the Influence of Exogenous Inputs and Latent Variables?" Journal of Neuroscience Methods 206 (1): 73–77.

[Rosenblatt 1958] Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model For Information Storage And Organization In The Brain. Psychological Review. 65 (6): 386-408.

[Rumelhart et al. 1986] Rumelhart, D., Hinton, G. and Williams, R. (1986). Learning representations by back-propagating errors. Nature.

[Russo et al. 2016] Russo, I., Frontini, F. and Quochi, V. (2016). OpeNER Sentiment Lexicon Italian - LMF, ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

[Salton 1975] Salton, G., Wong, A. and Yang, C. (1975). A Vector Space Model for Automatic Indexing. Communications of the ACM 18 (11): 613.

[Salton and McGill 1983] Salton, G. and McGill, M. (1983). Introduction to Modern Information Retrieval. McGraw-Hill Book Co., New York.

[Samuelson 1965] Samuelson, P. (1965). Proof that properly anticipated prices fluctuate randomly. Industrial Management Review 6, pages 41-49.

[Sander 2014] Sander, D. (2014). Recommending music on Spotify with deep learning. http://benanne.github.io/2014/08/05/spotify-cnns.html.

[Sarlin 2013] Sarlin, P. (2013). On policymakers' loss functions and the evaluation of early warning systems. Economics Letters, 119(1):1–7.

[Saxe et al. 2013] Saxe, A., McClelland, J. and Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In ICLR.

[Schindler et al. 2007] Hlaváčková-Schindler, K., Palus, M., Vejmelka, M. and Bhattacharya, J. (2007). "Causality detection based on information-theoretic approaches in time series analysis". Physics Reports. 441 (1): 1–46.

[Schmidhuber 2015] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks, 61:85–117.

[Schreiber et al. 2000] Schreiber, T. (2000). "Measuring Information Transfer". Physical Review Letters. 85 (2): 461–464.

[Schuster and Paliwal 1997] Schuster, M. and Paliwal, K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11):2673–2681.

[Segoviano and Goodhart 2009] Segoviano, M. and Goodhart, C. (2009). Banking Stability Measures. IMF Working Paper No. 09, 1-54.

[Semeniuta et al. 2016] Semeniuta, S., Severyn, A. and Barth, E. (2016). Recurrent dropout without memory loss. arXiv preprint arXiv:1603.05118.

[Sengupta et al. 2008] Sengupta, R. and Tam, Y. (2008). The LIBOR-OIS spread as a summary indicator. Economic Synopses - St. Louis Fed.

[Sheldon and Maurer 1998] Sheldon, G. and Maurer, M. (1998). Interbank Lending and Systemic Risk: An Empirical Analysis for Switzerland, Swiss Journal of Economics and Statistics (SJES), 134, issue IV, pages 685-704.

[Sims 1972] Sims, C. (1972). Money, Income and Causality. American Economic Review, 62(4):540–552.

[Socher et al. 2011] Socher, R., Pennington, J., Huang, E., Ng, A. and Manning, C. (2011). Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[Socher et al. 2013] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A. and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1631–1642. Association for Computational Linguistics.

[Soo 2013] Soo, C. (2013). Quantifying animal spirits: news media and sentiment in the housing market. Ross School of Business Paper No. 1200.

[Sprenger and Welpe 2010] Sprenger T. and Welpe I. (2010). Tweets and trades: The information content of stock microblogs. Available at SSRN: https://ssrn.com/abstract=1702854 or http://dx.doi.org/10.2139/ssrn.1702854.

[Srivastava et al. 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1):1929–1958

[Stevens 2012] Stevens, K., Kegelmeyer, P., Andrzejewski, D. and Buttler, D. (2012). Exploring Topic Coherence over Many Models and Many Topics. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '12, pages 952–961.

[Sutskever et al. 2013] Sutskever, I., Martens, J., Dahl, G. and Hinton, G. (2013). Proceedings of the 30th International Conference on Machine Learning, PMLR 28(3):1139-1147.

[Sutskever et al. 2014] Sutskever, I., Vinyals, O. and Le, Q. (2014). Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems (NIPS 2014).

[Swiss National Bank 2011] Swiss National Bank (2011). Swiss National Bank sets minimum exchange rate at CHF 1.20 per Euro. Press release.

[Tumasjann et al. 2010] Tumasjann, T., Sprenger, T., Sandner, P. and Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. Association for the Advancement of Artificial Intelligence, 1:178–186.

[Tumasjann et al. 2012] Tumasjann, T., Sprenger, T., Sandner, P. and Welpe, I. (2012). Natural Language Processing to the Rescue?: Extracting Situational Awareness Tweets during mass emergency Association for the Advancement of Artificial Intelligence, 1:178–186.

[Turian et al. 2010] Turian, J., Ratinov, L. and Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In ACL, pages 384–394.

[Turney and Pantel 2010] Turney, P.D. and Pantel, P. (2010) From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research, 37, 141-188.

[Upper and Worms 2004] Upper, C. and Worms, A. (2004). Estimating bilateral exposures in the German interbank market: Is there a danger of contagion?. European Economic Review, 48, issue 4, pages 827-849.

[Vasicek 1984] Vasicek O. (1984). Credit valuation. KMV corporation, March.

[Vicente et al. 2011] Vicente, R., Wibral, M., Lindner, M. and Pipa, G. (2011). Transfer entropy - a model-free measure of effective connectivity for the neurosciences. Journal of Computational Neuroscience, 30(1), 45–67.

[Wang et al. 2015] Wang, P., Qian, Y., Soong, F., He, L. and Zhao, H. (2015). A unified tagging solution: Bidirectional lstm recurrent neural network with word embedding. arXiv preprint arXiv:1511.00215.

[Werbos 1974] Werbos, P. (1974). Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. Ph.D. thesis, Harvard University.

[Whittaker 1990] Whittaker, J. (1990). Graphical models in applied multivariate statistics. Wiley Publishing.

[Wierstra and Schmidhuber 2007] Wierstra, D. and Schmidhuber, J. (2007). Policy gradient critics. Machine Learning: ECML 2007, pages 466–477.

[Witten 2011] Witten, W., Friedman, J. and Simon, N. (2011). New insights and faster computations for the graphical lasso. Journal of Computational and Graphical Statistics, Volume 20, Number 4, pages 892-900.

[Worldbank 2018] World Bank (2018). Exports of goods and services (% of GDP). World Bank national accounts data, and OECD National Accounts data files.

[Wu 2009] Wu, H. (2009). Global stability analysis of a general class of discontinuous neural networks with linear growth activation functions. Information Sciences, 179(19):3432–3441.

[Wyner 1978] Wyner A. (1978). A definition of conditional mutual information for arbitrary ensembles. Information and Control. 38 (1): 51–59. doi:10.1016/s0019-9958(78)90026-8.

[Yang et al. 2016] Yang, Z., Salakhutdinov, R. and Cohen, W. (2016). Multi-Task Cross-Lingual Sequence Tagging from Scratch. arXiv preprint arXiv:1603.06270.