Università
Ca'Foscari
Venezia

Master's Degree programme
in Economics and Finance

Final Thesis

Analysing the impact of E.C.B. Communication on Financial Markets

A Text Mining approach

Supervisors
Ch. Prof. Paolo Pellizzari
Ch. Prof. Enrico Maria Cervellati

Assistant Supervisor
Ch. Prof. Antonella Basso

Graduand
Alessio Venturini
Matriculation number
877631

Academic Year
2019 / 2020

Contents

List of Figures

List of Tables

Acknowledgements

The completion of my Master Thesis represents a significant milestone in my life. While challenging, the writing of this Thesis has been a fruitful, fascinating, and rewarding process. My experience would, of course, not have been the same without the support and friendship of several people. As such, I would like to thank them.

First, I am really grateful to my two Supervisors, Paolo Pellizzari and Enrico Maria Cervellati, for their support throughout the entire process. Regarding the former, I want to thank him for his ongoing support about the text mining topic, central to the realization of this Thesis. Regarding the latter, beside his help in this Thesis about Behavioural Finance insights, I want to thank him with respect the experience related to the CFA Research Challenge. For their trust in me, I am forever grateful.

Second, I am very thankful to my girlfriend Veronica who I met during my Bachelor studies at the Roma Tre University. To her, I have nothing but gratitude, and, as in the acknowledgements in my Bachelor Thesis, I renew my hope to continue in reaching together ambitious goals in our life.

I am also grateful to different people from the financial industry that helped me during these academic years, allowing me to assess financial events (also relevant for this Thesis) with a professional point of view. The people I am talking about are Egidio Nigro, Simone Curti (my mentor at Mentors4u) and Lorenzo Rigamonti (my mentor during the CFA Research Challenge). The former, in particular, gave me support since the first year of Bachelor degree, and our conversations and debates have generated several ideas relevant to this Thesis and for future research projects during my Ph. D. studies.

Additionally, I extend my gratitude to my family, that always believed in me in the last years. In particular, they have made enjoyable the low and high moments of this entire journey.

Last but not least, I am thankful friends for their tremendous support and friendship throughout my life. In particular, I want am grateful to Emanuele Lopetuso, “travelling companion” since the first year of high school, and with whom I will share my Ph. D. studies at the University of Reading.

Preface

This Thesis investigates how European Central Bank (ECB) communication made during the press conferences days affects asset pricing and herding behaviour in the European financial market.

My approach to analysing the ECB statement is through the lens of *text mining* analysis. Specifically, ECB press conferences are extracted by means of a Web Scraping algorithm [203] that divides ECB's President transcript from the answers in the Q&A section. Then, Sentiment Analysis [252] is applied to both these two types of ECB text corpus and used to construct a sentiment time series using different field-specific dictionaries, that span from the Monetary Policy [261] to Financial stability [76] lexicon. I also created my own dictionary merging different lexicons found from previous studies [76], [207]. These time series are then used as explanatory variables to compare to the Euro Stoxx 50 time series, considered to be a proxy of the European stock market returns [27], [52]. Evidence suggests that my sentiment time series reached the highest explanatory power in predicting Euro Stoxx 50 market realizations. To prevent my results against the possibility of spurious (time series) regression, I applied the Philip-Ouriss cointegration test [259], finding that mine and the Euro Stoxx 50 series are cointegrated. The relevance of this result could have important implications both in the economic and financial domain.

Additionally, to assess the evidence of whether herding behaviour of investors occurs around positive and negative spikes of the sentiment time series, I applied an augmented version of Chiang and Zeng [66] regression to different European sectors, finding that those spikes do help in predicting herding behaviour in the financial market. Notably, results vary across different types of sectors, a finding in line with previous studies [249].

My work is innovative for at least two kinds of reasons. First, I introduced a new dictionary to measure ECB press conferences sentiment and its linkage with the European stock market, according to the topical changes [182] detected in ECB communication and its Institutional (and supervisory) role during time [92]. More importantly, I investigated for the possibility of a long run relationship between my sentiment time series and the Euro Stoxx 50. Second, given the fact that literature on herding behaviour, event studies and reactions to central bank announcements has evolved independently [31], it seemed to be worthwhile to relate these themes. I then attempted to explain investor herding patterns across several European financial sectors using both positive and negative values in my sentiment time series as further control variables in the Chiang and Zeng [66] regression. To the best of my knowledge, this is the first work that attempts to detect herding in the European stock market focussing on the *direct* impact of ECB communication.

The Thesis is organized as follows. Chapter 1 provides a theoretical foundation of the *text mining* subject [77], describing in detail the difference between this field and the one related to *data mining* [50], as well as the text mining process that will be applied to ECB (text) data. It also provides an empirical example of how such a process can be used to investigate Twitter data related to the #COVID19 topic. Chapter 2 discusses text mining applications in a financial and central bank domain and provides a literature review useful for the last chapter of this Thesis. Chapter 3 describes the pattern of herding behaviour in the financial market, with a particular emphasis of its (possible) sources with respect to central bank policies. Chapter 4 concludes, describing how the text mining procedure was applied to quantify the impact that 123 ECB press conferences had on the European stock market during the period that spans from the 01/2008 to 10/2019.

Chapter 1

Text Mining

Text mining is a big and interdisciplinary field, and benefits from contribution of several correlated disciplines, namely information retrieval, data mining, machine learning, probability, statistics, and computational linguistics [77].

In this chapter, I will provide an introduction behind the potential of this subject, describing its main features and the related challenges in computational tractability. In particular, in section 1.1 we will frame the text mining concept in a Data Science prospect, while in section 1.2 its relationship with Machine Learning will be discussed. Subsequently, we will see text mining history, techniques, and process in sections 1.3, 1.4 and 1.5, respectively. In conclusion, in section 1.6 we will link all the material described in this chapter focussing our attention to the tools provided in the `R programming language`.

1.1 Text mining: A Data Science introduction

Data science is the subject that links statistics, data analysis, machine learning and their related methods in order to understand and analyse actual phenomena with data [140].

Data has been pivotal in human society ever since time immemorial and it has been written in the form of *text* to keep records of information that can be analysed in the future. Actually, while text data was mainly captured in the *physical* forms, with the advent of technology *digital* forms took its place[1] and therefore new frontiers for data science are possible.

Hence, in order to understand more in depth the relationship between data science and text data, we will now describe the different types of data we can encounter in data science problems, and the relative type of analysis that arises from their usage.

1.1.1 Types of Data

One of the first thing to understand in Data science applications is that not all data are *created* equal.

In general, we can identify at least four types of data[2]: (i) structured data; (ii) unstructured data; (iii) semi-structure data; and (iv) metadata. Each one varies for peculiarity about its storage, abundance on the World Wide Web (hereafter Web), eases in computational tractability and models we can apply to them.

[1] `https://analyticstraining.com/text-mining-the-next-big-thing-in-data-science/`

[2] `https://www.bigdataframework.org/data-types-structured-vs-unstructured-data/`

1. Structured data

When we talk about *structured* (or rectangular) data we mean collections of values that are each associated with a variable and an observation [25]. Common examples of structured data are *Excel* files or *SQL* databases. Structured data depends on the existence of a data model (i.e., a model of how data can be stored, processed and accessed): because of this model, each field is discrete and can be accesses separately or jointly along with data from other fields. Structured data are considered the most "traditional" form of data storage, since the earliest versions of database management systems (DBMS) were able to store, process and access structured data [270].

2. Unstructured data

For *unstructured* data we mean (applying a negative definition) datasets that do not naturally fit the paradigm of structure data: that is, they do not have a predefined data model or they are not organised in a pre-defined manner. Examples of such type of data are: *images*, *sounds*, *trees*, and ***text*** [335]. The ability to analyse unstructured data is especially relevant in the context of Big Data, and according to an important report from Edureka[3] a large part of data in organisations is unstructured, as we can see in Fig.1.1.

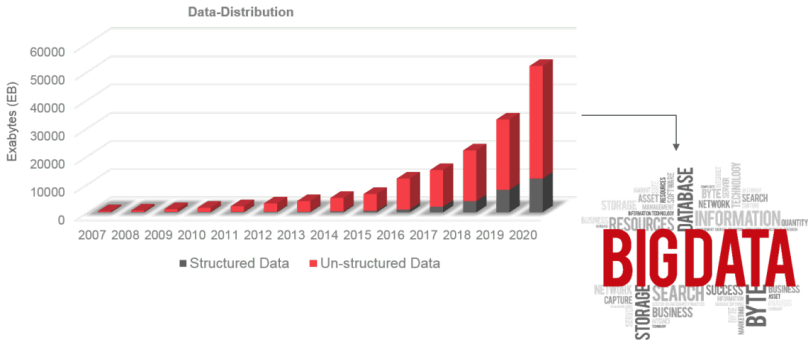


Figure 1.1: On-line data format distribution forecast by 2020.

The implication of this graph is that the ability to extract value from unstructured data is (and will be in the future) one of main drivers behind the quick growth of Big Data.

In this Thesis, the only unstructured data we will focus our attention are textual data.

3. Semi-structured data

In the middle between the two first types of data we find the so-called *semi-structured data*: they are a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contain tags (or other markers) to separate semantic elements and enforce hierarchies of records and fields within the data [2]. Examples of semi-structured data include JavaScript Object Notation (hereafter JSON), HyperText Markup Language (also known as HTML) and eXtensible Markup Language (hereafter XML).

[3] `https://www.edureka.co/blog/10-reasons-to-learn-hadoop/growth-of-unstructured-data-learn-hadoopedureka/`

4. Metadata

The last category of data type is metadata. From a technical point of view, this is not a separate data structure, but it is one of the most important elements for Big Data analysis [348]. Metadata is "data about data". They provide additional information about a specific set of data. For example, in a set of photographs, metadata could describe when and where the photos were taken. The metadata thus provides fields for dates and locations which, by themselves, can be considered structured data.

1.1.2 Types of Data Analysis

The distinction provided so far is important when we talk about the difference between *data mining* and *text mining*. Broadly speaking, we can define data mining as the practice to find hidden patterns in large datasets [50], whilst text mining as the practice to see hidden patterns inside (a large set of documents containing) text. Even though data mining and text mining are often seen as complementary analytic processes that solve problems through data analysis, they differ on different issues[4].

Firstly, while data mining handles *structured* data, text mining deals with *unstructured* (textual) data.

Secondly, on one hand, data mining combines disciplines including Statistics, Artificial Intelligence and Machine Learning to be applied directly to structured data. On the other hand, text mining requires an extra step while maintaining the *same* analytic goal as data mining: in particular, before *any* data modelling or pattern recognition function can be applied, the unstructured (text) data has to be organized and *structured* in a way that allows for data modelling and analytics to occur. This is a really important point and will be analysed in more details in section 1.5.

For the moment, linking what we said so far, we can define text mining as [84]: "*large-scale, automated processing of plain text language in digital form to extract data that is converted into useful* ***quantitative*** *or* ***qualitative*** *information*".

1.2 Machine Learning and Text mining

Machine learning (ML) is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed [7].

In general, the information discovered by data mining and text mining can be learnt by ML algorithms and applied to new data. What is important is that the patterns found are useful to explain the data and/or make prediction from it.

In this Thesis the focus will be on the ML algorithms that can be applied for text mining applications. In particular, in this section we will briefly describe some of these algorithms, and we will show how they can be applied to different (text) decision problems in section 1.4.

ML algorithms can be classified into *supervised* or *unsupervised* learning depending on the goal of the algorithm.

In the first case, ML usually involves a subsample of data that is used to fit the algorithms, and another subsample used to test its predictive accuracy. The former subsample of data is known as the "*training*" data and the latter is the "*test*" data

[4] `https://blogs.opentext.com/whats-the-difference-between-data-mining-and-text-mining/`

(out of sample). There may even be a third hold-out sample, known as "*validation*" data, used for final testing of models that have passed muster on test data.

In the second case, instead, we do not have a training set but only the test set: as we will see, this both represent an advantage (since unsupervised algorithms can be applied to a much wider range of problems) but also a limit (since we do not have a metric to measure their performance).

1.2.1 Supervised Algorithms for text mining

In general, when a number of explanatory X variables are used to determine (or *predict*) some outcome Y (and we train an algorithm to do this) we are performing supervised ML [300]. The variable Y is referred to as *response* or *output* variable. The output variables are in general assumed to be dependent on the inputs. Hence, the goals of supervised algorithms are to uncover the relationship between input/output variables and to generalize this function so it can be applied to new data.

Supervised learning algorithms can be further divided into *classification* and *regression* algorithms.

1. Classification Algorithms

In the case of **classification** ML algorithms, the output variable, also called *class* (*qualitative*, *categorical* or *factor*) variable, is a factor taking two or more discrete values in a finite set [185]. Instead, the input variables can take either continuous or discrete values. The goal is to predict qualitative or categorical outputs which assume values in a finite set of classes, by means of a *classifier* [258].

A full description of all classification ML algorithms is beyond the scope of this Thesis, but here we will briefly describe the most important ones for text mining purposes, and later we will see their application in section 1.4: the algorithms we are talking about are the Naïve Bayes, Decision Tree and the Support Vector Machine.

- **Naïve Bayes**
  In general, *Naïve Bayes* algorithm is a classification technique based on the well-known Bayes' theorem, reported here:

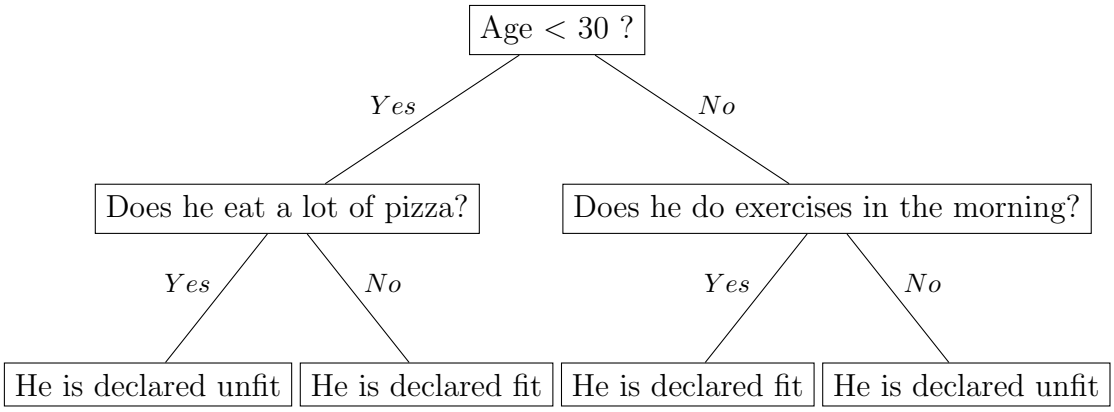
$$P(\theta|x) = \frac{P(\theta)P(x|\theta)}{P(x)}. \tag{1.1}$$

  Where:

  - $P(\theta|x)$ is the posterior probability of class given its predictor (or attribute),
  - $P(\theta)$ is the prior probability class,
  - $P(x|\theta)$ is the likelihood which is the probability of predictor given class,
  - $P(x)$ is the prior probability of predictor.

  It is important to notice that this algorithm assumes *independence* among predictors [349]. This assumption implies that the presence of a particular feature in a class is unrelated to the presence of any other feature. Despite its "simplicity" at first glance, Naïve Bayes is known to outperform even highly sophisticated classification methods in several datasets [60].

- **Decision Trees**
  A decision tree algorithm is an algorithm where each node represents a feature (or attribute), each link (or branch) represents a decision, and each leaf represents an outcome [284]. In general, the graphical representation allows for a better comprehension of how this algorithm works, as we can see from the following classification task between "fit" and "unfit" people:



  The instances are classified going down from the root to the leaf nodes: in text mining applications the nodes in the tree are simply *text*.
  From this graph we also understand that the tree construction of a tree may be seen as a *variable selection method*: at each node the problem is to select the variable to divide upon and how to perform the split. To avoid over-fitting (that is, too many nodes to classify instances) several methods can be applied. The most famous are the *pre-pruning* [49], that stop growing the tree earlier before it perfectly classifies the training set, and the *post pruning* [244], that instead allows the tree to perfectly classify the training set and then post prune the tree.

- **Support Vector Machines** (SVM)
  In the case of Support Vector Machines (SVM), an instance is viewed as a p-dimensional vector, and the idea is to discover if it is possible to separate such instances with a $(p-1)$ dimensional hyperplane.
  Formally, the goal of the SVM is to map a set of entities with inputs $X = \{x_1, x_2, ..., x_n\}$ of dimension n, i.e., $X \in R^n$, into a set of categories $Y = \{y_1, y_2, ..., y_n\}$ of dimension m, such that the n-dimensional X-space is divided using hyperplanes, which result in the maximal separations between classes Y.
  A hyperplane is the set of points $\mathbf{x}$ satisfying the equation $\mathbf{w} \cdot \mathbf{x} = b$, where b is a scalar constant, and $w \in R^n$ is the normal vector to the hyperplane, i.e., the vector at right angles to the plane. The distance between this hyperplane and $\mathbf{w} \cdot \mathbf{x} = 0$ is given by $b/||\mathbf{w}||$, where $||\mathbf{w}||$ is the norm of vector $\mathbf{w}$.
  In general, SVM algorithm tends to provide better in sample and out of sample results respect to the aforesaid classification algorithms [1], hence their widespread popularity.

2. Regression Algorithms

In an opposite way respect to classification algorithms, the goal of **regression** ML algorithms is to predict a real-valued number for the output variable [147]. The dependency of Y on X can be "described" using a function called regression function, or $E[Y|X]$. Also here, we will only describe the regression algorithms that are relevant for text mining application, and in general the most important one (in particular for text classification tasks, as we will see in in section 1.4) is the *logistic regression* [112].

Logistic regression is the most common regression for *binary* classification and belongs to the *generalized linear model* models [98]. In classification task (and under a mathematical point of view), when the output variable is dichotomous the logistic regression is more suitable than the *linear* one: the fact is that the underlying probability distribution of Y is a Bernoulli and not a Gaussian one. Moreover, in a logistic regression problem we are not interested in the predicted value for Y, but instead we search for the *probability* that a certain element belongs to one of the two classes.

Hence, this regression produces a probability that an event could happen with the following equation:

$$logit(y) = c_0 + c_1x_1 + c_2x_2 + ... + c_kx_k \tag{1.2}$$

where $logit(y)$ is the logit function and $y = \frac{\exp(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k}{1+\exp(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)}$ is the expected probability that the outcome is present.

To conclude, we have to notice that the logistic regression uses a more complex *cost function* than the one used from linear regression. In particular, this cost function can be defined as the "*Sigmoid* function" or also known as the "*logistic* function". The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore, linear functions fail to represent it as it can have a value greater than 1 or less than 0, which is not possible as per the hypothesis of logistic regression [145].

1.2.2 Unsupervised Algorithms for text mining

Unsupervised learning is applied when we only have X variables and no separate outcome variable Y to predict or classify [120]. These methods are used to identify key features of the data and to explain its structure by grouping instances more alike or by transforming the data in smaller dimensions.

Among the different tools provided from unsupervised ML algorithms, the ones of interest in text mining applications are *clustering* and *topic modelling*.

1. Clustering algorithms

Clustering is used to group instances that seem to share some similarity [344]. The goal is to find *clusters* and to assigns instances to them. The success of clustering algorithms is often measured *subjectively* in terms of how useful the result appears to be to a *human* user, since as we said at the beginning of this section there is *no* measure of performance (as we have in supervised learning). Effectiveness of the algorithm will be thus a matter of opinion and cannot be verify with objectivity, but the advantage is that these kinds of algorithms is that they can be applied to a broader set of problems with respect to the ones that can be addressed from supervised ML algorithms.

The most famous clustering technique is the *k-means* algorithm: the objective is to divide a set of observations (that can be a set of documents, for example) into K clusters by assigning them to the cluster with the closest centre [201]. The idea behind this algorithm is that clusters will be defined by K centroids, where each centroid is a point that represents the centre of a cluster. This algorithm works interactively, where initially each centroid is placed randomly in the vector space of the dataset and move themselves to the centre of the points which are closer to them. In each new iteration, a centre is found for each cluster by minimizing a dissimilarity measure (distance) across all observations. The algorithm will conclude these iterations when the position or the groups don't change anymore.

2. Topic modelling algorithms

Topic modelling is an unsupervised ML technique that allows one to investigate the thematic features that are embodied in a set of unlabelled documents *topics* [245]. In particular, these models detect word and phrase patterns within documents, and automatically cluster word groups and similar expressions into a set of *topics*.

Among others, *Latent Dirichlet allocation* (LDA) is a particularly popular method for fitting a topic model [42]. The LDA model is guided from two important principles:

- *Every document is a mixture of topics.*
  We can say that each document may contain words from several topics in particular proportions. For example, in a two-topic model we could say "Document 1 is 85% topic A and 15% topic B, while Document 2 is 40% topic A and 60% topic B."

- *Every topic is a mixture of words.*
  For instance, we could imagine a two-topic model of Italian news, with one topic for "politics" and one for "entertainment." The most common words in the politics topic might be "President", "Parliament", and "Premier", while the entertainment topic may be made up of words such as "football", "film", and "actor". Importantly, words can be shared between topics: a word like "budget" might appear in both equally.

LDA is a mathematical method for estimating both of these two principles at the same time: it allows us to find the mixture of words that is associated with each topic, and to determine the mixture of topics that describes each document [297].

1.3 Text Mining History

Once we have understood that textual data belong to the unstructured type form and the different ML techniques an analyst has at hand, it is now interesting to step back and retrace the history behind the analysis of textual data. A deep understanding of the history of text mining is important for three reasons [235]:

1. To provide the context in which text mining developed,
2. To show the development paths followed in text mining techniques,
3. To avoid making the mistakes of the past.

Interest in text mining applications has grown in earnest since Turing's publication [325]. In particular, in his seminal work, Turing laid out his criterion for "intelligence": a computer could be considered intelligent if it can interact with humans without them ever realizing they were dealing with a machine. Text mining at its core is one of the embodiments of that vision where analysts obtain useful insights from textual data without ever needing to know whether they are interacting with a machine.

Rapidly, text mining relevance spread also in the business field. In particular, H.P. Luhn [212] was one of the first researcher to understand the power of text (and more in general, unstructured) data as means of enterprise information. Hence, in the 1958 he introduced the concept of *Business Intelligence* (BI) [212], defining it as a system that will: "*[...] utilize data-processing machines for auto-abstracting and auto-encoding of documents and for creating interest profiles for each of the "action points" in an organization. Both incoming and internally generated documents are automatically abstracted, characterized by a word pattern, and sent automatically to appropriate action points*". The important point here is to notice that the earliest BI focus was on text rather than on numerical data. Moreover, Luhn was the first (or among the firsts) to develop many of basic techniques that are now common in information science and that we will see in more detail in section 1.4. These techniques included full-text processing, self-indexing, hash codes, automatic abstraction and the concept of selective information dissemination.

Some decades later, Don R. Swanson invited scientists to be more like "intelligence analysts", and to take seriously the idea that new knowledge should be gained from document collections, stating that [313] : "*[...] New knowledge [...] is seen as emerging from large numbers of individually unimportant but carefully hoarded fragments that were not necessarily recognized as related to one another at the time they were acquired. [...] The analyst is continually interacting with units of stored data as though they were pieces selected from a thousand scrambled jigsaw puzzles.* ***Relevant patterns, not relevant documents, are sought***". This definition implicitly describes that new knowledge derives from a large number of documents that would not be really relevant if considered individually and not necessarily recognized as linked to each other when they were acquired. Swanson was implicitly inviting analysts to consider the concept of *Corpus*, an argument that we will explain in section 1.5. It is in fact Swanson that put in practice this idea in the Biomedical literature, creating the Arrowsmith software [302], which allows the user to find interesting co-occurrences thanks to common keywords and phrases "*complementary and non-interactive*".

In the same years, important (and metaphorical) definitions were used for text mining, such as: *"mining implies the extraction of precious nuggets of minerals from worthless rock*", by Hearst [142] or "*the gold hidden in [...] mountains of textual data*" [99]. Notably, Hearst did not just attempt to explain what text mining is, but compared it with the concept of "*information retrieval*" (IR). He specified that IR mainly deals with the extraction of important documents, leaving the user to fully understand the semantic content. Instead, in text mining applications we try to derive or discover new knowledge from data.

It is now clear that when we want to use a text mining in data science problems, we should try to respect the following characteristics:

- It must operate on collections of large texts written in a *natural language*,
- It has to use *algorithms* rather than manual or heuristic filters,
- It must generate *new knowledge*.

1.4 Text Mining Areas

Actually, text mining is in a loosely organized set of competing technologies that works as analytical "city-states" with no clear dominance among them. According to the one of the most important books about text mining [235], we can relate these technologies to seven different areas. The unifying theme behind each of these techniques is the need to "turn text into numbers" so that ML algorithms (listed in section 1.2) can be applied to large document databases. Though distinct, these areas are highly interrelated and often a typical text mining project will require different techniques from several of them.

The seven practice areas will be described in each of the following paragraphs.

1.4.1 Information Retrieval

Search and *information retrieval* (IR) covers indexing, searching, and retrieving documents from large text databases with keyword *queries*[5] [217].

Document's ranking is strictly related both to the concept of IR and query: formally, given a query q and a collection D of documents that match the query, the problem is to rank (that is, to sort) the documents in D according to some criterion so that the "best" results can be displayed to the user.

This field increased in importance with the advent of the internet, and the Google's PageRank algorithm [277] developed in 1998 (which is still nowadays a key part of Google's method of ranking web pages in search results), is clearly an empirical example.

1.4.2 Natural Language Processing

Natural language processing (NLP) refers to methods and techniques allowing a *computer* to directly handle text in human language [188]: the general goal is to infer the underlying structure of text at different possible levels of detail. For natural language, we mean any language that we use in everyday life (such as English, Italian, Chinese, etc.) as opposed to the *formal* and *computer* language.

In general, NLP contribution to text mining is on providing linguistic data (e.g. documents' annotations, part-of-speech tags, parsing results) on the information extraction phase.

In the following, some common NLP tasks are presented.

Part-of-Speech Tagging

In general, each word in a text belongs to one part of speech (POS), a grammatical category denoting its function in a sentence [289]. Most common parts of speech are *nouns* and *verbs*, usually presented in every phrase, while examples of other parts are *articles*, *pronouns* and *conjunctions*.

While for a human is often trivial to distinguish among such different parts of speech, for computers is not. For example, the word ***set*** might be: (i) a noun ("the *set* of natural numbers"); (ii) a verb ("parameters must be *set*"); or (iii) an adjective ("the table is *set* for two people").

[5] In Informatics, a query is the interrogation of a database in order to extract from it the data that satisfy a certain criterion.

Specialized algorithms for POS tagging exist, analysing sentences in their entirety to determine the correct POS for each of their words: known examples of these tagged datasets, which are also used for evaluating the correctness of the methods, are the Brown Corpus [114] and the Penn Treebank Dataset [314].

Sentence Boundary Detection

Sentence Boundary Detection (or *sentence breaking* or *sentence segmentation*) is the problem in NLP of deciding where sentences "begin and end" [273]. For certain languages this process is (quite) simple: for English, Italian, German and other types of western languages, it is almost as easy as finding every occurrence of punctuation like " . " ; " ? " ; or " ! " in the text. However, also if some problem can arises (in particular in cases where there could be periods that occur as part of abbreviations or acronyms), most of times few simple heuristic rules that can correctly identify the majority of sentence boundaries are used without "sacrificing" important information [68].

On the other hand, for some specific language these rules do not hold: in the case of Arabic, Chinese and Japan for instance, there are no blank spaces that divide words, and other approaches have been developed for these special cases [308].

Lemmatization

Lemmatization is the process of grouping together the different inflected forms of a word so that they can be analysed as a single item [320]. For example, "computers" is an inflected form of "computer", the same logic as "singing" being an inflected form of "sing".

Lemmatization is used in NLP and many other fields that deal with linguistics in general. Moreover, it provides a productive way to generate generic keywords for search engines or labels for concept maps [263].

1.4.3 Text Classification and ML Classification

In general, *text classification* (also known as text *categorization*, or *topic spotting*) is the process of automatically sorting a set of documents into categories from a predefined set [292]. Text classification is often the first step in the selection of a set of documents to submit to further processing, or it can be the only step in text processing.

A well-known example of text categorization is *spam filtering*: a spam filter examines an incoming email and determines whether to mark it as spam or let the message pass untouched into the in-box. This is a binary classification task, since only two outcomes (or class labels) are possible: "spam" and "not spam". Hence, the basic approach to text classification is to derive a set of features to describe a document, and then apply an *algorithm* designed to select these features in order to classify that document in the appropriated category.

The (supervised) algorithms described in section 1.2.1 are widely used for text classification tasks.

It is important to notice that in recent years, more advanced techniques have been proposed to implement classifiers able to tackle tasks in support of document's ranking, a concept introduced in section 1.4.1. In particular, *hierarchical* classification (where text can be naturally categorized into hierarchical topic trees, where branches in the tree

indicate finer-grained categorical distinctions) is preferred to *flat* classification (where each document is classified without taking into account any hierarchical structure) and improvements are documented in different papers [121], [345], [312].

1.4.4 Text Clustering and relative ML algorithms

Text clustering uses unsupervised algorithms to automatically group similar documents or words into clusters [159]. Whether this kind of clustering is used to group documents in a corpus or words in a document, the technique is almost always used as a means to an end, rather than the end itself.

Despite the growth in interest in topic modelling (a concept already described in 1.2.2), one of the most important subfield of text clustering is *text similarity*.

Text similarity has to determine how "close" two pieces of text are both in surface closeness (lexical similarity) and meaning (semantic similarity) [123]. In general, the idea is to represent documents as vectors of features and compare documents by measuring the distance between these features.

There are multiple algorithms and different metrics to compute sentence similarity [158], but here we will only see the *cosine similarity* [269], defined as follows:

$$\cos(\mathbf{t}, \mathbf{e}) = \frac{\mathbf{te}}{\|\mathbf{t}\|\|\mathbf{e}\|} = \frac{\sum_{i=1}^{n} \mathbf{t}_i \mathbf{e}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{t}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{e}_i)^2}} \tag{1.3}$$

The cosine similarity provides a solution to evaluate the angle between two vectors without considering their length: the smaller is the angle, the more similar are the documents. To obtain a similarity measure constrained in the range between 0 and 1, the cosine of such angle is considered, which is equal to 1 for a null angle and is non-negative if all the vectors are so.

The cosine similarity is advantageous because even if two similar documents are far apart by the *Euclidean distance* (due to the size of the document), chances are they may still be oriented closer together[6].

1.4.5 Concept Extraction

Concept extraction is the technique of mining the most important topic of a document. [352]. Because documents are typically a loosely structured sequence of words and other symbols (rather than concepts), the problem is not trivial, but it can provide powerful insights into the meaning, provenance and similarity of documents.

In text mining applications, the analyst could be interested in finding the *sentiment* from a set of documents, or the general *relationship between words* in these documents.

Sentiment Analysis

Sentiment Analysis (or *opinion mining*) is the interpretation and classification of polarity (positive, negative) within text, whether a whole document, paragraph, sentence, or clause [252]. It is important to notice that actually exist techniques that allow for a more fine-grained sentiment classification, and permit to identify specific emotion an author is expressing (like fear, joy or anger) [305].

Roots of sentiment analysis are in the studies on public opinion analysis at the beginning of 20^{th} century and in the text subjectivity analysis performed by the

[6] `https://medium.com/@adriensieg/text-similarities-da019229c894`

computational linguistics community in 1990's [220]. However, the outbreak of computer-based sentiment analysis only occurred with the availability of subjective texts on the Web and now its usage spans from the analysis of social media up to the analysis of financial markets, as we will see in more details in chapter 2.

In this Thesis, Sentiment Analysis will be applied to ECB Press Conferences in order to understand if they can affect the European financial market.

Word Association and Text Networks

Recently, the study of the representation of concepts and word associations has gained interest. In particular, a number of studies have explored the distributional and structural properties of word association *networks* showing that the structure of connections between associations adheres to special topological laws, commonly found in many natural networks such as small-would properties [87].

Network analysis refers to a family of methods that describe relationships between units of analysis. A network is comprised of *nodes* as well as the *edges* (or connections) between them [44]. Though network analysis is most often used to describe relationships between people, novel approaches use networks in order to find relationships between words. For example, one can represent a corpus of documents (a concept that we will see in section 1.5) as a network where each node is a document, and the thickness or strength of the edges between them describes similarities between the words used in any two documents. There are multiple advantages to use a network-based approach to automated text analysis. Regarding what we said in the previous section, network representation for textual data represents an arguably more sophisticated technique for identifying clusters within text data.

1.4.6 Information Extraction

Information Extraction refers to the automatic extraction of structured and semi-structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources [79]. The goal of IE is to extract *only* the semantic information relevant in the text, that can be useful for the human user and easy to manipulate for a computer.

Among others, *Named Entity Recognition* (NER) is one of the most important field when we talk about IE: NER is an extraction process that searches to extract and classify each element in the text in predefined categories, such as people, organizations, currencies, etc [223]. The idea behind a NER is the so-called knowledge base, that contains a list of known concepts (or Named Entities) that can be extracted from the text of interest.

Applications of NER can be seen both in academic area (in order to improve the research efficiency [264]) and industrial one (mainly used for customer care [311]).

1.4.7 Web Mining

Due to the unique structure and enormous volume of data appearing on the web, we can say that *Web mining* is actually its own practice area [75]: it is not much different from text mining in written documents, but a relatively unstructured mining and analytic field has developed around the web documents, so web mining has developed in a rather confined niche. In particular, as the Internet becomes even more related with our popular culture thanks to the rise of Facebook, Twitter, and other social

media channels, web mining will continue to increase in value and will be an important "ally" of text mining. This is related to the fact that the volume of unstructured data companies will have at hand will increase more and more, as showed in Fig. 1.1.

Web mining can be divided into three different types: (i) web *usage* mining; (ii) web *content* mining; (iii) and web *structure* mining. In this Thesis, the focus will be on the latter. *Web content mining* mainly deals with the extraction of useful information from the contents of web pages [203]. By content, we mean all the unstructured data listed in section 1.1, i.e. images, videos, links, audio, and *text*. As we will see, scraping content from webpages involves a series of tasks. First, one must identify the web pages that need to be scraped or the *regular expressions* (that is, a sequence of characters that define a search pattern [29]) that would match with the *Uniform Resource Locator* (or URLs) that need to be scraped. Next, we need to identify the data-points that we need to capture. Once these things are done, we will have to set up the infrastructure in a certain programming language in order to mine content from the web[7].

As we can see in Fig. 1.2 the concept of text mining is vast and embodies connection between all the different fields described so far. In particular, an instructive image in this regard is provided from [235], which shows these linkages by means of a Venn diagram:

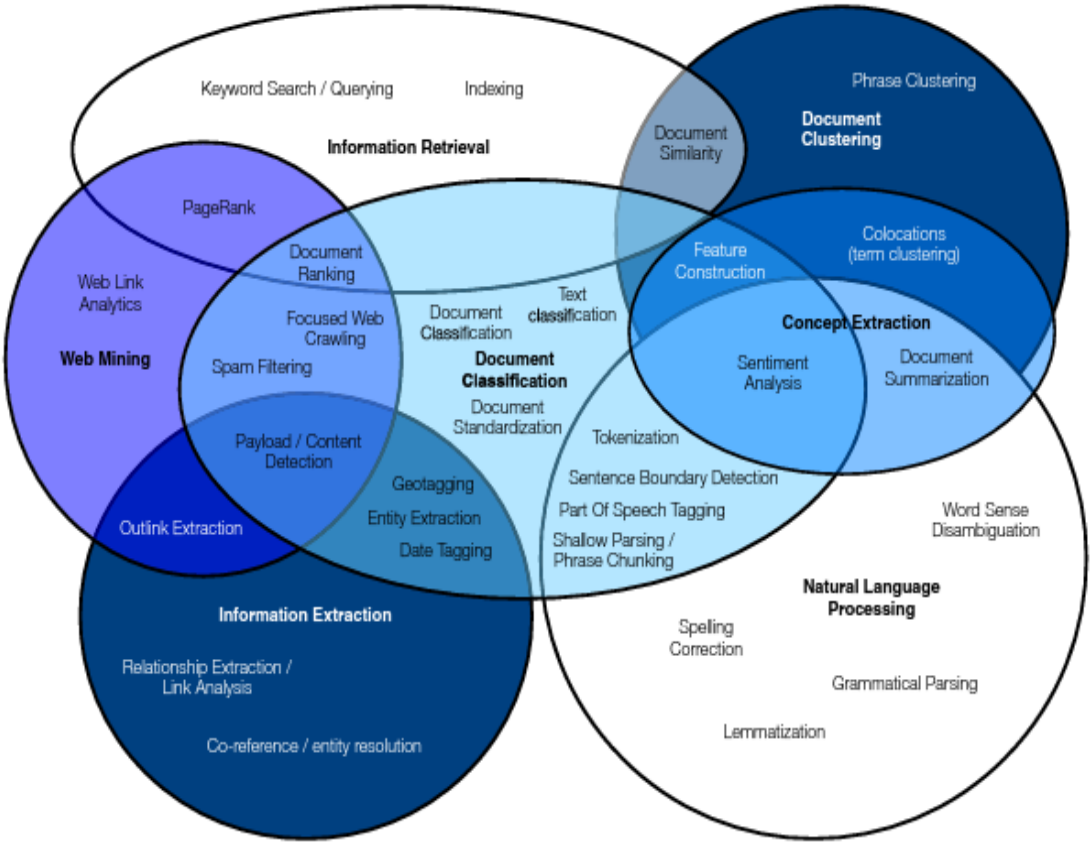


Figure 1.2: Interactions between text mining areas [235].

[7] `https://medium.com/@promptcloud/web-content-mining-the-approach-application\-and-future-7deabeef928a`

1.5 Text Mining Process

Before to apply one of the different areas (to our textual documents) listed section 1.4, we have to transform text in something a "computer can understand" [281]. More formally, given the unstructured data type form which text data belong to, we have to manipulate our documents in order to transform them in a structured form: the final goal is to apply models to these data in order to extract new useful patterns.

Hence, in each of the following paragraphs I will describe all the different phases behind text mining analysis.

1.5.1 Download of textual data and creation of a Corpus

The first step (in every Data Science problem) is the retrieval of data. In text mining applications, the choose of textual-type data is huge, since possible sources can be: Web pages, Twitter posts, e-mail, books, newspapers and so on. It is important to notice that these data could also have different formats, such as HTML, .pdf, .txt, JSON. For sure, each format will require its own difficulty in computational tractability.

Given the question we want to give an answer, the different set of documents accumulated from the analyst will create a *corpus*: a corpus simply represents the union of all these documents (irrespective of format), each providing an additional piece of information to the argument of interest [228]. For instance, in order to analyse the way in which a certain politician talks to the public, we will create a set of text files with her speeches: this set of text files will represent our corpus to answer to the different questions of interest.

1.5.2 Structured representation of textual data

In general, there are different methods in order to structure textual data according to the objective of our analysis. We will now present two different types of text data representation, but more emphasis will be given to the former due to its wider usage.

1. The vector-space model, weighting schemes and n-grams

The most popular structured representation of text is the *vector-space model*, which represents text documents as a vector whose elements indicate the occurrence of words within text [286]. For sure, this results in an extremely high-dimensional space since every distinct string of characters occurring in the collection of text documents has a dimension. Importantly, this model makes an implicit assumption (called the *bag-of-words* assumption) which implies that any information about the order or structure of words in the document is discarded [351]. In other words, the model is only concerned with *whether* known words occur in the document, not *where* in the document: the intuition is that documents are similar if they have similar content[8].

Graphically, what we will obtain is a matrix where each row corresponds to one of the q documents in our corpus, and the columns with the relative n words d that we can find in them. We will call this matrix *document-term-matrix* (or DTM) (also if sometimes also a term-document matrix, or TDM, can be constructed) and its graphical representation can be seen in table 1.1[9]:

[8] `machinelearningmastery.com/gentle-introduction-bag-words-model/`

[9] The numbers of terms are pure representatives.

Document	d_1	d_2	...	d_n
$Document_1$	1	4	$\cdots$	2
$Document_2$	0	0	$\cdots$	0
$Document_3$	3	1	$\cdots$	1
$\cdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$Document_q$	0	2	$\cdots$	0

Table 1.1: *A Document-Term-Matrix*

Also if this such a representation of documents could be too restrictive, for many text mining tasks (such as document classification or clustering) it seems to be reasonable: in particular, the collection of words appearing in the document is usually sufficient to differentiate between semantic concepts and to develop different ML algorithms described in section 1.2.

In conclusion, it is important to notice that in *this* kind of representation we are making two implicit assumptions.

The first one is that each word enters in the matrix according to its (absolute) *frequency* (or more simply, term frequency, or $t_f(d, q)$) across all documents. Anyway, other ways of computing these values (also known as *term weights*) have been developed: well-known examples are the (i) *inverse document frequency* [218] (or simply *idf*); (ii) the *term frequency-inverse document frequency* [341] (hereafter *tf-idf*); and (iii) the (more recent) Word2vec method [234], belonging to the family of Word embedding. Anyway, we will now describe formally only the first two weighting schemes, since the latter doesn't seem to have found applications in Economics[10]:

- The *idf* weighting coefficient
  Quite intuitively, articles that contain words that are uncommon across all documents are more likely to be similar to other articles that contain those words, whilst articles that have common joint words are less likely to be similar [218]. This is the idea behind the *idf* weighting coefficient, where, formally, the *idf* for word d would be:

$$w_d^{idf} = \log\left(\frac{N}{n_t}\right) = -\log\left(\frac{n_t}{N}\right) \tag{1.4}$$

  where N is the total number of documents, and n_t is the number of documents containing word d.

- The tf-*idf* weighting coefficient
  Following a similar logic with respect to the *idf* coefficient, the idea behind the tf-*idf* score is that common terms that appear in most documents earn low scores because they are less informative (i.e. they have low *idf*), as do terms that are rare in a given article (they have low $t_f(d, q)$). Formally, he tf-*idf* for word d would be:

$$w_d^{tf-idf} = t_f(d, q) \cdot w_j^{idf} \tag{1.5}$$

[10] `https://cbail.github.io/textasdata/word2vec/rmarkdown/word2vec.html#word2vec-models-with-kerasneural-networks`

where $t_f(d, q)$ is the number of times that term d occurs in document q, and w_d^{idf} is the weighting scheme defined in eq.(1.4). Hence, The *tf-idf* transformation defines the most representative terms in a given document to be those that appear infrequently overall, but frequently in that specific document [118].

The second assumption is that we are considering only *one* word (or *uni-gram*) per document. Also here, a different way to pursue the analysis is to use *n-grams* (generally from two up to five words together). In some situation, such a strategy could be more valuable for the entire analysis [96], [261].

2. Sequential representation

Though the bag-of-words assumption works well in many cases, it is not a universal solution. For some tasks, such as automatic translation of languages and NLP, the order of words is critical for solving the task successfully.

Hence, for more complex and specific text mining problems, specialized algorithms and models for handling sequences such as the *finite state machines* [157], *conditional random fields* [227], or *sequential pattern algorithm* [226], are used.

In this Thesis we will work with the first type of representation for textual data, but it was worth noting the existence of the ones just described, in particular for future text mining applications.

1.5.3 Preprocessing

When we talk about (text) *preprocessing* we mean all the different techniques that allow us to convert unstructured and semi-structured text into the structured vector-space model, in order to analyse our data in a quantitative and qualitative way [328].

It general, preprocessing is the most important phase of the whole process and most of time (and efforts) spent on the analysis can be attributable to this part (from here we can also understand why the sentence "80% of data science is really data cleaning[11]" could not be any truer).

The preprocessing steps we will describe now are the same for all text mining tasks (though which processing steps are chosen depends on the task), and involve: (i) tokenization; (ii) removal of stopwords; (iii) stemming; (iv) removal of capital letters, punctuation and numbers. In chapter 4 of this Thesis, such a process will be applied to structure ECB Press conferences data.

1. Tokenization

With tokenization, we mean the reduction of a corpus into a vector of words [225]. This technique divide text contained in the corpus in tokens (i.e. atomic block of text, typically made up of indivisible characters). A *token* can be defined as any sequence of characters surrounded by delimiters.

In general, the main problem is to find these delimiters. A first approach could be the one that see at spaces, tabulations and new line characters as the only delimiters needed. Anyway, the problem here is that in each text (and also according to a different language) there are a lot of exceptions, as we described in section 1.4.2. It is thus crucial to utilize tools (such as the Sentence Boundary Detection algorithms)

[11] `https://www.infoworld.com/article/3228245/the-80-20-data-science-dilemma.html`

to create tokens that are coherent with the text we are analysing and the relative language in which this text is written.

2. Removal of stopwords

"Stopwords" are non-contextual words, i.e., not germane to interpretation of the text that are removed from the data before conducting textual analysis [339]. For instance, some stopwords such as "*the*", "*are*", "*as*" are simple to see as being redundant, but others may be more subtle.

The advantage of this procedure is to reduce the quantity of words to extract from texts, thus lowering the computational efforts. Moreover, this will result also in an increase of the overall quality of our data.

To ease the stopword process, during years different vocabularies have been developed, both for English language (such as the one from Stanford NLP [219]) and for other languages of interest, such as Chinese [354].

3. Stemming

With stemming we mean the reduction of any inflected form of words and verbs to their root [208]: for example, "I have", "she has" are both two forms that can be traced back to the verb "to have". The creation of a stemming algorithm has been one of the most challenging problem in Informatics, and different algorithms were provided [172]. They are different according on their accuracy, performance and in the way in which resolve the required task. As for the previous step, stemming allows us to have cleaner analysis and to focus only on the word's root.

Lemmatization (presented earlier in section 1.4.2 as subfield of NLP) is similar to word stemming but it does not require to produce a stem of the word but to replace the suffix of a word, appearing in free text, with a (typically) different word suffix to get the normalized word form. Anyway, given the deeper knowledge required to create the dictionaries that allow the algorithm to look for the proper form of the word, in general a stemmer algorithm is preferred to a lemmatizer algorithm[12].

Lastly, also here different algorithms vocabularies have been developed for each language.

4. Removal of capital letters, punctuation and numbers

In conclusion, other cleaning procedure are possible, including the removal of capital letters, punctuation and numbers. Anyway, this phase could change according to the goals of the analysis: for instance, if our analysis requires to identify firms' names (such as in the case of NER, as we saw in section 1.4.6) it would not have sense to remove capital letters. From this we can understand why sometimes the application of this (or one of the steps listed before) step is conditioned to the peculiarity of our text, or to the goal of our research.

1.5.4 Application of text mining models

In general, in order to find the right model to apply to our text data the logic provided in Fig. 1.3 below could be of interest [235]: as we can see, group of questions must be

[12] `https://blog.bitext.com/what-is-the-difference-between-stemming-and-lemmatization/`

answered to determine the appropriate processing task and to drive the analyst to the choice of the best model to pursue her analysis.

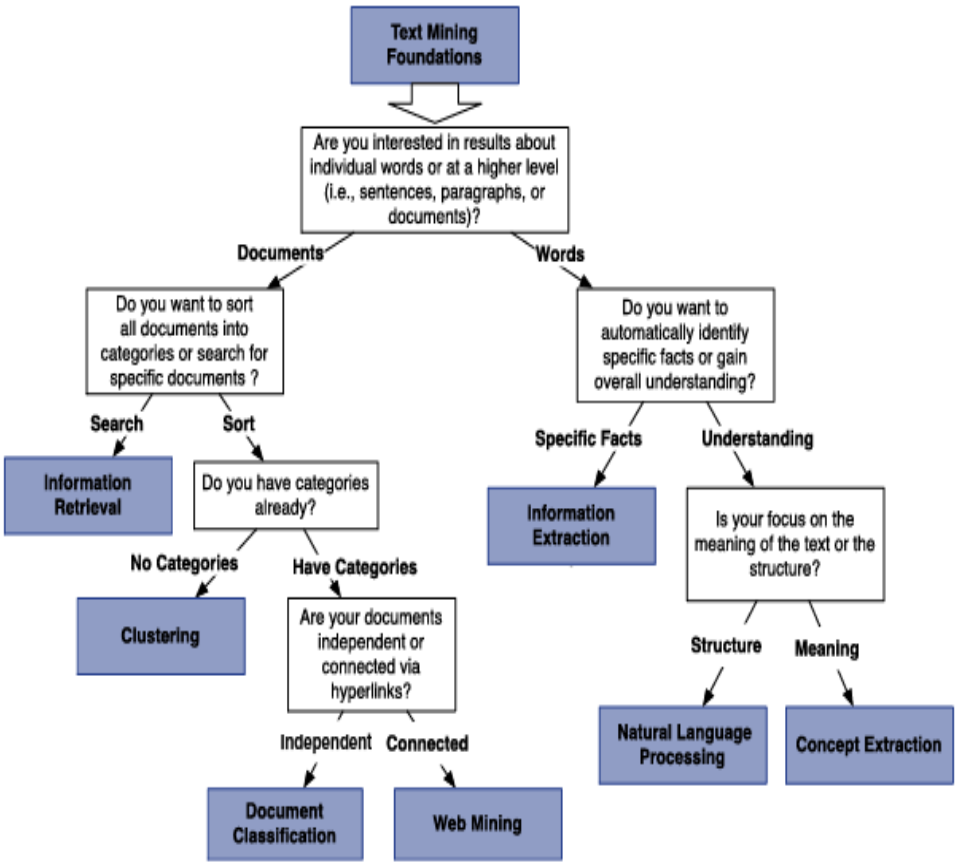


Figure 1.3: The path to decide our text mining model [235].

At the same moment, it is worth noting that rarely will a single pass through the tree solve any text mining problem. A text mining solution usually consists of multiple passes through the data at different levels of processing [235].

1.5.5 Visualization of obtained results

Visually representing the content of a text document is one of the most important tasks in the field of text mining[13].

The most used types of graphs when we talk about text mining are the *term frequency plot*, *wordcloud*, *cluster* and *network* representation. We will see some of these graphs in the next section, entirely dedicated to text mining implementation in the `R programming language`.

1.6 Text Mining in R Programming Language

In general, `R` is one of the languages of choice for text mining (and Data science) but is not the only one. Other language processes exist for the same purposes, both open source (such as Python and Java), and commercial ones (such as SAS Text Miner,

[13] `https://towardsdatascience.com/a-complete-exploratory-data-analysis-and\protect\discretionary{\char\hyphenchar\font}{}{}visualization-for-text-data-29fb1b96fb6a`

IBM® SPSS® Modeler Text Analytics, etc.). Anyway, the main advantage of `R` is that it is a full-fledged Econometrics environment, and extracted textual data may be instantly visualized, quantified, and analysed statistically.

In the following, we will see: (i) the two main data text format types for text mining applications; (ii) an implementation of text mining process in R environment; and (iii) the most important packages that an analyst has at hand for text mining tasks.

1.6.1 Types of text formats in R and Text Mining

When we work with `R` it is important to understand that we can follow two different approaches to manage text formats: *tidy* and *"not-tidy"*.

The principles of *tidy data* provide a standard way to organize data values within a dataset [333]. In general, tidy data has a specific structure: (i) each variable is a column; (ii) each observation is a row; (iii) each type of observational unit is a table. We thus define the tidy text format as being *a table with one-token-per-row* [297]. The resulting tidy data sets allow manipulation with the standard set of "tidy" tools with a lot of packages that we will present in section 1.6.3.

At the same time, most of the existing R tools for NLP (besides the `tidyverse` package) are not compatible with this format, and the CRAN Task View for Natural Language Processing[14] lists a large selection of packages that take other structures of input and provide *non-tidy outputs*. In particular, these packages store or manipulate text in one of the following formats:

1. ***String***
   In this case text will be stored within `R` as character vectors (and often text data is first read into memory in this form).

2. ***Corpus***
   These types of objects typically contain raw strings annotated with additional metadata and details.

3. ***Document-term-matrix***
   This is a sparse matrix describing a collection (i.e., a corpus) of documents with one row for each document and one column for each term. The value in the matrix is typically word count or *tf-idf* as we stated earlier in section 1.5.

Anyway, as we can see from Fig.1.4 (as presented in [297]) it is possible to *connect* the tidy text format with other important packages and data structures, allowing the analyst to rely on both existing text mining packages and the suite of tidy tools.

In this vein, in this Thesis I will show how an analyst can use both the tidy and non-tidy data text formats and relative tools for text mining applications.

1.6.2 Text mining process in R: An empirical application

It is now interesting to present a practical case to understand which are the tools provided in `R` to apply (in practice) the text mining process described in section 1.5, thus knowing what is happening "under the hood" of the software.

The case of interest presented here refers to the analysis of Twitter data about the #COVID19 topic.

[14] `https://cran.r-project.org/web/views/NaturalLanguageProcessing.html`

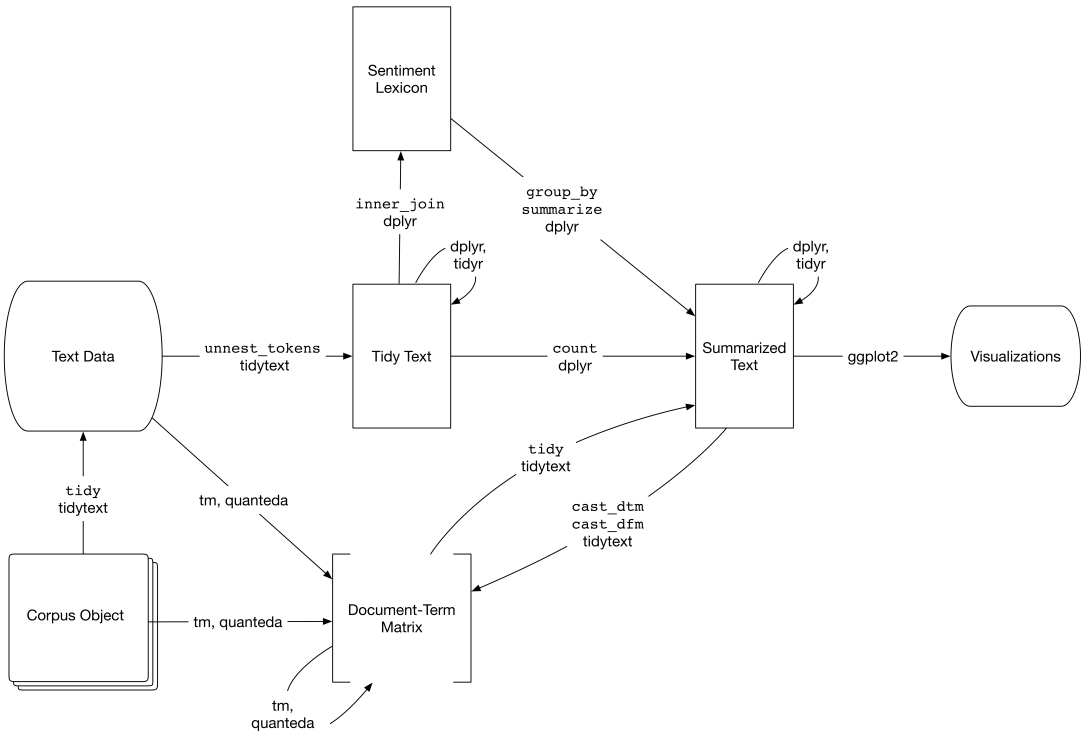


Figure 1.4: The relationship between tidy and not-tidy text applications [297]

1. Download of textual data

In order to download data from Twitter, an account for Twitter developers must be created at the link `developer.twitter.com`. This account is needed because it will allow us to create the key *application programming interface* (or API) needed to download data from the site. Actually, the main packages to work with Twitter data are `twitteR` and `rtweet`, also if recently more attentions is paid to the former since the latter does not receive updates from its developer anymore.

Both packages provide a function to access to the Twitter's API: the first via `setup.twitter.oauth()` and the second via `create.token()`. The various parameters required in both functions (such as the consumer key and consumer secrets) can be found at the same link above. Then, in order to download data from the site also here two different functions are provided.

Regarding the first package, we can use the `searchTwitter()` function, where the first argument refers to the topic of interest, the second to the number of tweets to extract (5,000 in this case), and the third to the language of interest.

Regarding the second package, we have to use the function `search-twitter()`, where the first two arguments are the same, but the third argument instead allows the user to extract the *most recent* or the *most popular* tweets about the argument (I chose the former option). It is important to notice that the output of this function will be stored in a `data.frame` format and will contain results about the date of creation of that tweet, its number of re-tweets, the language (not only English) and other kinds of information.

The difference in the implementation of these two functions can be seen in the code below:

```
bd1 <- searchTwitter("#covid19", n = 5000, lang = "en")
bd2 <- sapply(bd1, function(x) x$getText())  # Extract text

df <- search_tweets("covid19", n = 5000, type = "recent")
```

In Fig.1.5 we can see the frequency (or volume) of tweets that we downloaded (as resulting on 6^{th} of May, 10:00 am). The spikes in the (time) series are not easy to analyse since we are using frequency in "seconds" and twitter posts come from different parts of the world.

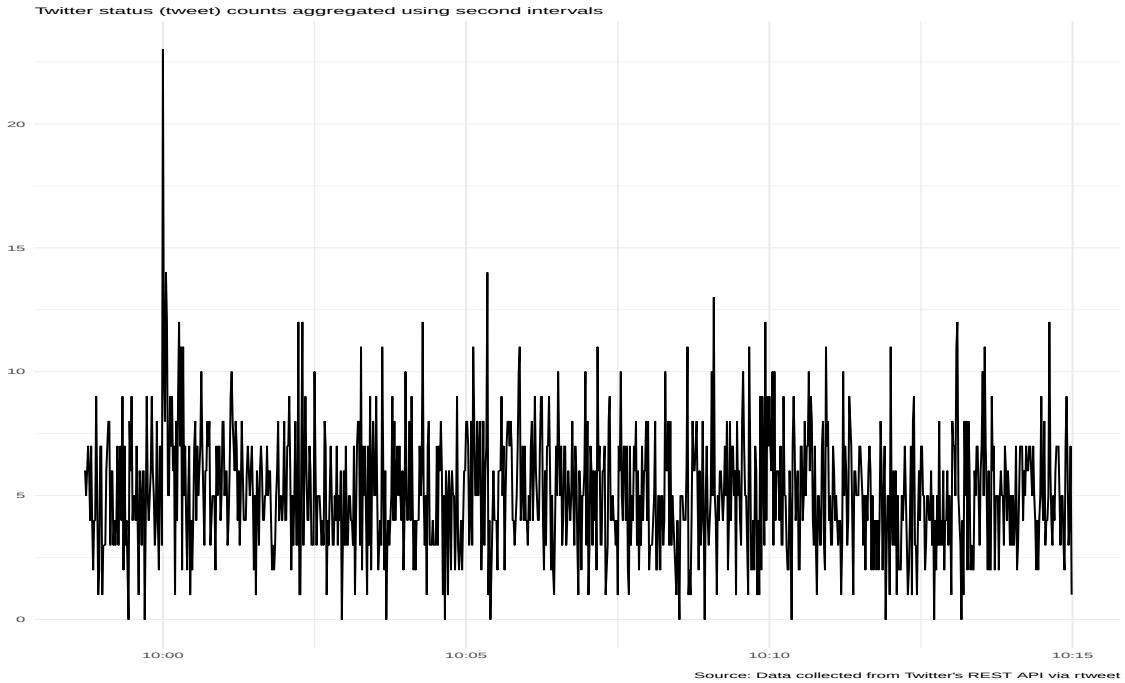


Figure 1.5: Volume of Twitter post about the #COVID19 topic

Hence, the following analysis will be pursue using only the `twitteR` package in order to focus only on *English* posts.

2. Structured representation of textual data and preprocessing

I decided to merge these two steps since both can be tackled via the `tm` package: in particular, a corpus for these twitter data can be created via the (auto-explicative) `Corpus` function. Then, the preprocessing phase is applied thanks to the `tm_map()` function, allowing the user to remove capital letter, punctuation and stopwords, via the arguments `content_transformer(tolower)`, `removePunctuation` and `stopwords()` respectively, as we can see from the code below:

```
bdc <- Corpus(VectorSource(bd2))

# We apply the preprocessing process
bdc <- tm_map(bdc, content_transformer(tolower))
bdc <- tm_map(bdc, removePunctuation)
bdc <- tm_map(bdc, function(x)removeWords(x, stopwords()))
```

To remove special characters that could lower the quality of our analysis, I used an ad hoc function that can be found in Appendix A.

Having cleaned (and structured) data at hand, we can create a TDM via the function `TermDocumentMatrix()` in which we will store uni-grams according to their *absolute* frequencies. From this matrix, useful information can be derived, such as the correlation (Pearson coefficient) between terms and a certain benchmark word (in our case "virus") thanks to the function `findAssocs()`. Also if we do not show here all the outputs, it is interesting to see that the aforesaid benchmark term has a correlation of 0.37 with *laboratory*, and it could be related to the fact that in these days America accused China of creating the Covid-19 virus into laboratories[15].

3. Models

In order to understand what could be the main "hot topic" about #COVID19, we could be interested in *cluster* twitter posts into specific sub-arguments. Formally, we can apply the k-means algorithm described in section 1.2.2. In `R`, we can use the (built-in) `kmeans()` function that requires the user to insert the number k of cluster required. Given the (not so huge) number of twitter data we have, we set $k = 3$.

This function will elaborate three different kinds of clusters, and in our case:

1. $cluster1 : covid, will, new$
2. $cluster2 : covid, coronavirus, people$
3. $cluster3 : \ldots, covid, lockdown$

As we said in section 1.2.2, effectiveness of the algorithm is a matter of opinion, and from here we can infer that these three clusters (should) refer to: (i) the possibility of new cases around the world; (ii) the relationship with Coronavirus family (to which the Covid-19 belongs); and (iii) the lock-down that is applied from governments all around the world (we have to notice that in $cluster3$ we see the presence of "..." due to the re-tweets posts).

4. Visualizations of results

In conclusion, it is interesting to see our results in a graphical way. As we said in section 1.5, two of the most famous representation for text data are the term frequency plot and the word-cloud, presented in Fig. 1.6 and 1.7, respectively.

Both these figures allow the analyst to understand which are the words that were more related with the #COVID19 topic. As one would expect, the words *people*, *lockdown* and *cases* tend to co-occur with this argument.

1.6.3 Important packages for text mining in R

In order to ease the text mining process, several packages have been developed in `R` during years. Now, we will see different packages according to the area of interest, that allow us to do a lot of operation described in this chapter and that we will use in the final chapter of this Thesis.

[15] `https://www.cnbc.com/2020/05/03/us-intelligence-documents-accuse-china-of-\covering-up-coronavirus-outbreak.html`

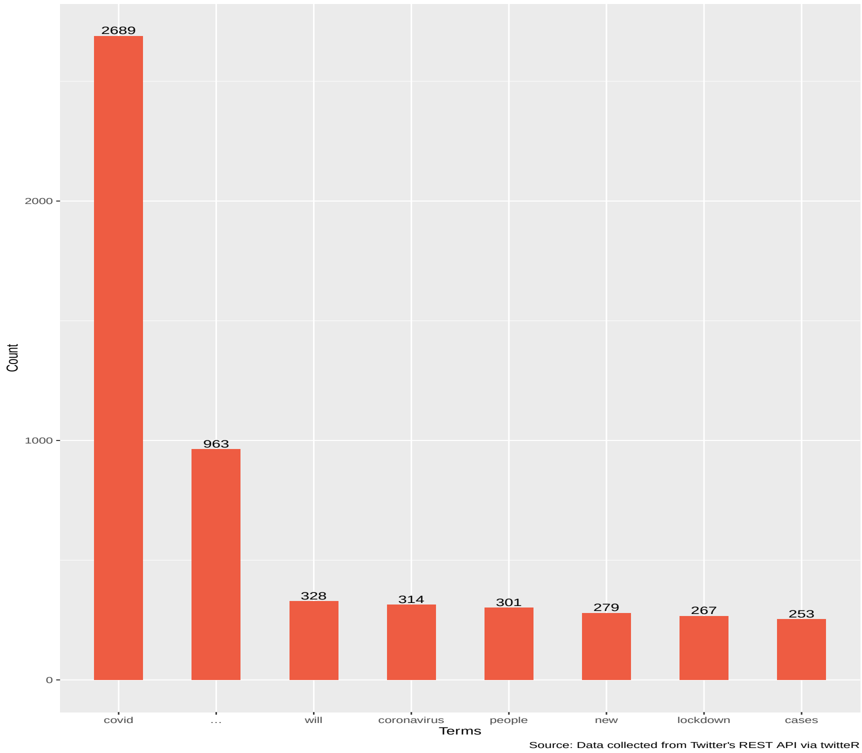


Figure 1.6: Most common terms related to the #COVID19 topic

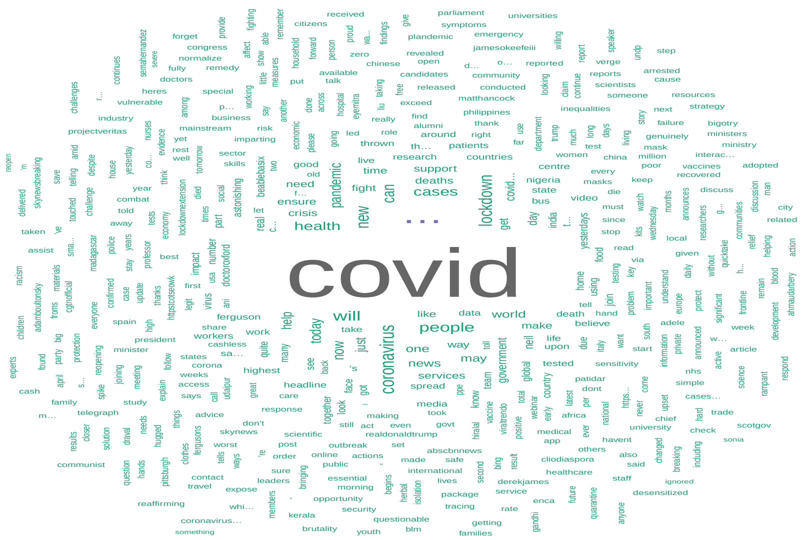


Figure 1.7: Wordcloud for Tweets related to the #COVID19 topic

For what concern the data import we can use the `dplyr` package (that in turn is inside the `tidyverse` [338] package cited earlier). For the Web (content) Scraping instead, we can use the `rvest` [336], `RSelenium` [137] and `XML` [192] packages.

Instead, `tidyverse` and (especially) `tm` [108] provides a list of useful functions to apply all the preprocessing phases (such as tokenization, stemming, stopwords, etc.).

For string manipulation, one of the most important package is `stringr` [337].

Text mining models can be found in several packages according to the application of interest. Machine Learning models related to both structured and unstructured data can be applied in `R` via the `e1071` [97], `caret` [187], `maptree` [332] and `rpart` [317] packages. The implementation of Neural Networks are provided in `keras` [13] and `TensorFlow` [150] packages. Instead, useful dictionaries and word lists (particularly relevant for Sentiment Analysis) are provided in the `sentometrics` [12], `SentimentAnalysis` [111], `qdap` [125] and `Syuzhet` [173] packages.

In conclusion, for what concern text data visualization, `wordcloud` [109], `ggplot2` [334], `plotrix` [199], and `igraph` [81] are really valuable for analysis, providing functions to deal both with tidy and not tidy data.

Chapter 2

On the current state of art of Text Mining in Economics

The words' relevance in an economic contest can be promptly understood by the two following sentences and their (related) effects on financial markets:

1. "*We want Facebook to be somewhere where you can start meaningful relationships*", Mark Zuckerberg said on 1 May, 2018.
   The announcement created gasps, not just from the crowd in front of whom Zuckerberg was talking, but also in *equity* markets[1]. The share price of Match Group (the company that owns Match.com, Tinder and other dating websites) plunged by more than 20%, as we can see in Fig.2.1a.

2. "*[...] we are not here to close spreads. This is not the function or the mission of the ECB*", Cristin Lagarde said on 12 March, 2020.
   This *unexpected* sentence created fear on financial markets. On the same day, the European stocks had their worst performance day in history[2]. In particular, Germany's DAX and France's CAC 40 both plunged over 12%. Europe's Stoxx 600 fell 11% (the index's worst day on record), as we can see in Fig.2.1b.

Why are these examples so important? The answer is simple: in both cases, financial markets were being affected by a *sentence* made up of just a few words. *There was* ***not a single number*** *in the announcements.*

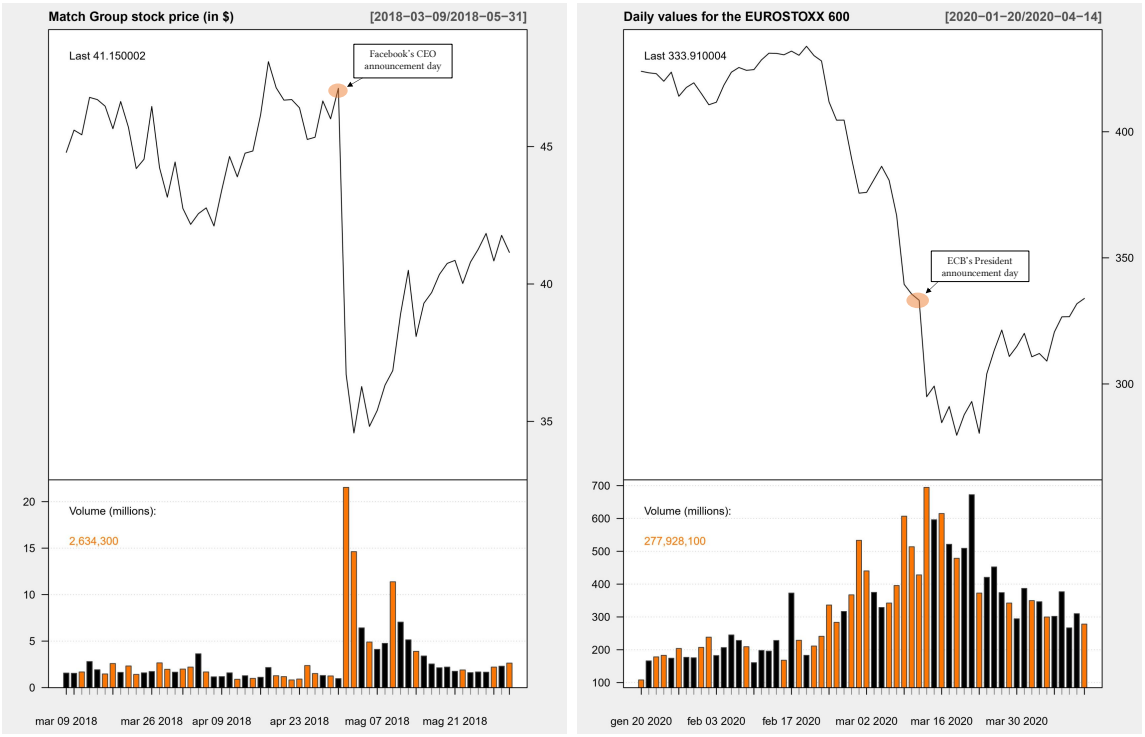
More interestingly, regarding the first case, Zuckerberg's comment did not impact Facebook's share price but the biggest effect was felt by Match.com, a company that until that moment may not have even been considered a competitor to Facebook. In the second case, instead, (regardless of the theoretical opinions that one could make about this event) few hours later the ECB's President tried to "calm" financial markets with an interview on CNBC[3]. It is interesting to notice that the content of this interview can be found in the Q&A section of the press conference of that day. According to the newspaper "Il Foglio" it is the first time ever the ECB inserts in a press conference an extract of an interview in order to clarify Presidents' speech[4].

[1] `https://www.man.com/maninstitute/shakespeare-without-the-monkeys`

[2] `https://edition.cnn.com/2020/03/12/investing/european-central-bank-coronavirus/index.html`

[3] `https://www.cnbc.com/video/2020/03/12/lagarde-we-will-absolutely-fight\-fragmentations-in-bond-markets.html`

[4] `https://twitter.com/davcarretta/status/1238388891104817153`



(a) Match Group stock market reaction to Facebook's announcement.

(b) EUROSTOXX 600 market reaction to ECB's announcement.

Figure 2.1: Two examples of how sentences can affect financial markets.

These events (i.e. few words causing strong reactions rippling through markets) happen frequently in financial markets, albeit usually more subtly. Moreover, in order to show how *market sentiment* could drive *herding* behaviour, I also added the relative chart about the volume for that particular security. As we can see, results clearly show a similar pattern. We will come back in more detail on this point in chapter 3.

Hence, it is straightforward to understand the different reasons for the gained interest in text mining applications described in Chapter 1 from both the economic and financial field, attracting academics as well as industrial practitioners. Regarding the former, in last years the publications on text mining applications increased exponentially [257]. Regarding the latter, different reports from the Chartered Financial Analyst (CFA) institute highlight the importance to integrate text mining tools in the financial industry [166], [194].

However, despite the strand of research of text mining application in Economics is flourishing, we will focus only on works relevant for the purpose of this Thesis. Specifically, in section 2.1 we will provide a literature review of the actual *state-of-art* in the *finance* field, whilst in section 2.2 we will see how text mining techniques are used in order to analyse *central banks' communications*. It will be interesting to notice that text mining used to predict events in financial markets is an interdisciplinary research that requires findings from linguistics, machine learning and behavioural finance, a point that is often highlighted in different papers [177], [241].

2.1 Text Mining in Finance: A General Review

In general, the quality of the interpretation of sentiment about the *sources* of text provided to financial practitioners each day can determine the *predictability* of assets'

value, *volatility* and *volume*, thus opening the floor for gains or losses [241].

According to Lu et al. [209], we can categorize these sources into three categories: (i) forums, blogs and wikis; (ii) newspaper and market research reports; (iii) content generated by firms and related (equity research) analysts that follow them.

It is worth stressing that the underlying goal of these sources has been (in most cases) *prediction* of market (or stocks') movement. This objective is not consistent with the efficient market hypothesis [215] (or EMH). Therefore, text mining algorithms that *can* predict market movements are empirical refutations of the EMH. Hence, in the following three sections, we will both see how these sources are used in general (rather than focusing on a specific subject) and, in *some* cases, as evidence to confute the EHM hypothesis. To ensure a capillary literature review of text mining in financial applications, different papers offering a systematic examination of past studies in this regard have been analysed: [6], [84], [118], [179], [189], [206], [237], [241], [257], [324].

2.1.1 Mining Forums and related textual content

Early work of text sources in the Finance domain focused on extracting sentiment and other information from messages posted to stock message boards, such as Yahoo! Finance, Motley Fool, Seeking Alpha, Raging Bull, etc. For instance, Wisocky [343] uses a sample of over $3,000$ stocks listed on Yahoo! Finance message boards, finding that message volume predicted next day abnormal stock returns. In particular, he finds that the total message posting volume is, on average, higher for firms with extreme past stock return and accounting performance, higher P/E and B/M ratios, higher past volatility and volume, higher analyst following and lower institutional holdings. However, some of these results are confuted by Tumarkin and Whitelaw [322] some year later, who examining the relationship between Internet message board activity (on RagingBull.com discussion forum), abnormal stock returns and trading volume, found that returns following abnormal Internet message board activity are statistically insignificant, a result consistent with EMH. The fact that message board postings do not predict market returns is confirmed from different works in the following years [9], [85]. Rather, it seems that this kind of text source can predict both volume and volatility in financial markets [9], [85].

Text mining also opens up the variety of sources of information. A peculiar example is provided from Bagnoli et al. [20], which compare the unofficial crowd-sourced forecasts of quarterly earnings from small investors posted to web pages (also called "whisper") to the First Call analyst forecasts. Their analysis shows that the former are more accurate than the latter thus finding that information about whisper is "impounded" in price prior to the earnings release.

In more recent times, however, a string of papers has attempted to exploit social media feeds, and in particular, those from the Twitter platform. As we said in section 1.6, this is due to the easiness API interface provided by Twitter and the vast amount of tweets produced each day available for the researcher. Notably, it seems that Twitter data seem to have some predictive power in stock returns. Bollen et al. [43] combine Twitter feeds and relate textual information to understand if public mood could have predictive power for the Dow Jones Industrial Average (DJIA). They found that it did: by constructing two different mood-tracking tools, namely the *OpinionFinder* that measures positive vs. negative mood and the *Google-Profile of Mood States* (GPOMS) that instead measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy), their results indicate an accuracy of 86.7% of the time in predicting the daily up and down changes in the closing values of the index. Some year later,

Rao and Srivastava [272] analyse more than 4 million tweets between June 2010 and July 2011 for the DJIA, NASDAQ-100 and 13 other big cap. They constructed a sentiment index by means of the Naive Bayes classifier (as described in section 1.2.1), showing that this index and returns have an 88% correlation. Moreover, Granger causality regressions indicate that tweets are predictive of stock and index movements. In conclusion, it is important to notice that tweeter data not only seem to forecast stock returns, but also stock volume and volatility [306], [323], [350].

2.1.2 Mining Newspapers and related textual content

Whereas forums and social media offer one approach to eliciting sentiment information, newspaper span a wider range of topics, allowing the analyst to enlarge her field of analysis beyond the concept of asset prediction (also if first works moved in this direction).

According to Gentzkow et al. [118] the first work about stock prediction using newspaper dates back even in 1933 [80]. In that work, Cowles tried to predict the future returns of the DJIA by (subjectively) classifying articles from the Wall Street Journal (hereafter WSJ), during the period 1902 - 1929 as "bullish", "bearish", or "doubtful". However, results do not seem to be impressive: a market-timing strategy WSJ editorials would had underperformed a passive investment in the Dow Jones Industrial Average by 3.5% per year. Nowadays, also if the implementation of stock prediction using text data is in general *computationally* driven, the logic applied to construct algorithms is similar to Cowles's approach. For instance, Leinweber and Sisk [198] compute sentiment from the Thomson Reuters NewsScope Engines, classifying news stories as bullish, neutral or bearish: $\{-1, 0, +1\}$. They show that their aggregate sentiment appears to track market conditions very well, as we can see in Fig 2.2, where the resultant time series drops after the financial crisis of the 2008.

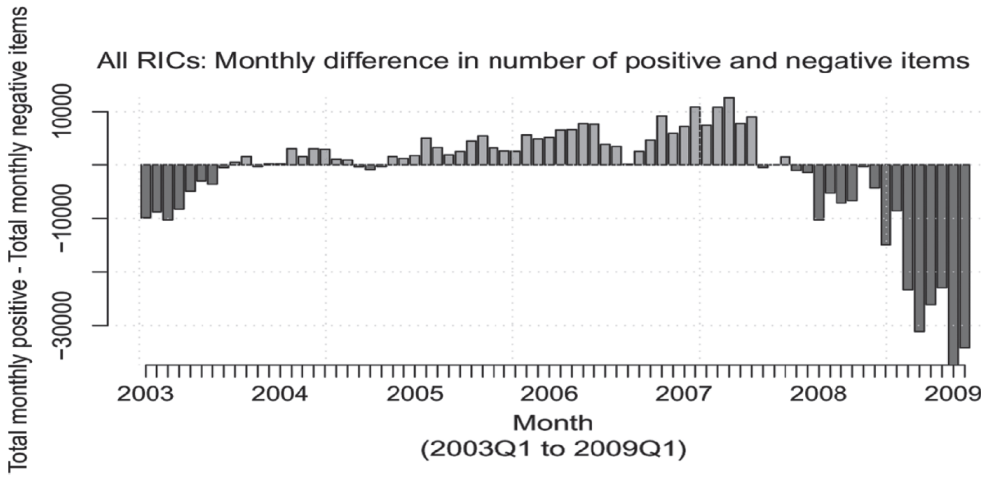


Figure 2.2: Reproduction from the *Journal of Portfolio Management* of Exhibit 7 from Leinweber and Sick [198], showing the net sentiment in aggregate from 2003-2009.

However, recently other (and more) sophisticated approaches have been proposed. Ming et al. [236] use a unified latent factor model to characterize the joint correlations between stock prices and news articles from the WSJ. They found that this model has an accuracy rate of 55.7%, quite higher respect to many widely used algorithms before. Kanungsukkasem and Leelanupab [177] introduced the concept of Financial Latent Dirichlet Allocation (FinLDA), an extension of the LDA model discussed in

section 1.2.2, with the peculiarity to take financial time series into account in order to improve feature extraction from text for prediction. They find that their features extracted from FinLDA empirically gave value added to prediction when used with Backpropagation Neural Network and Support Vector Regression algorithms.

As we said early, the usage of newspapers allows for an investigation of a wide list of topics and their direct (or indirect) effects on financial markets. Following this idea, Engle et al. [102] construct a time series (showed in Fig. 2.3) that captures news about long-run ***climate risk*** by calculating the *cosine similarity* (introduced in equation 1.3) between the *tf-idf* scores (explained in equation 1.5) for each daily text content of WSJ and a climate change vocabulary. Yet, Bybee et al. [55] extract the full text content of 800,000 WSJ articles for the 1984-2017 period, estimating a topic model in order to characterize the topical structure in business news. The most interesting result is that, since almost all topics exhibit strong time series persistence, this could be seen as an intuitive explanation behind ***volatility clustering*** events frequently observed in financial markets. For instance, they found that five-topic (i.e. "Recession", "VIX", "Electronics", "Problems" and "Small Business") news attention regression captures 63% of the variance in stock market volatility, as we can see in Fig. 2.4. Regarding ***policy uncertainty***, Baker et al. [23] develop the well-known index of economic policy uncertainty (or EPU), based on newspaper coverage frequency. Their first work constructed the index focussing only on the articles in ten leading US newspapers that contain the following triple: "economic" or "economy"; "uncertain" or "uncertainty". After their contribution, a new strand of empirical works gained attention, applying the same logic in other countries, like for instance in Belgium [319] or in Turkey [285]. Recently, the same researchers developed the Global Economic Policy Uncertainty (or GEPU) index and its graphical representation (here reproduces in in Fig.2.5) can be seen on their website[5]. As one can notice, the uncertainty scenario due to the Covid-19 impact all around the world is clearly the cause behind the spike in the last part of the time series [24].

2.1.3 Mining Company and Equity Research Reports

According to different papers [257], [324] the main datasets used for textual analysis have been company and analysts reports. The reason can be promptly understood: the quality of these text sources is much more technical than in message posting or newspapers, so one should expect richer meaning extraction. Text mining approaches in this domain try to address questions that are more in the *Corporate Finance* and *Risk Management* space.

It is also important to notice that, during years, textual analysis in this area has resulted in technical improvements, and "rudimentary" approaches such as word count methods have been extended to weighted schemes (introduced in section 1.5.2), where weights are determined in statistical ways.

Despite the "simplicity" of the former case, nowadays we can still find different papers using this approach, and often with significant results. For example, Nagar and Schoenfeld [240] introduced a climate risk measure at firm level by computing the frequency of the term "*weather*" in 100,000 firms' annual reports during the period 1994-2018. One of the most interesting finding is that whilst in 1994 the 25% of firms mentioned the term weather in their annual report, in 2018 this percentage increased

[5] `https://www.policyuncertainty.com/index.html`

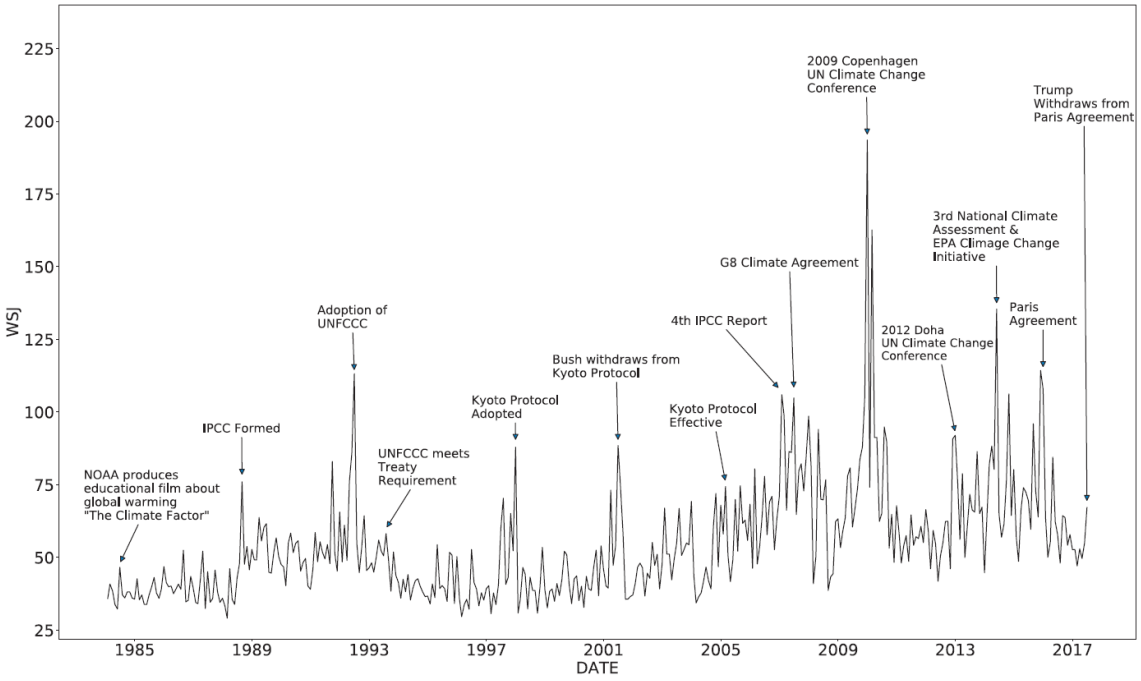


Figure 2.3: Reproduction from the *The Review of Financial Studies* of Exhibit 2 from Engle et al. [102], showing the WSJ Climate Change News Index during the period 1984-2017.

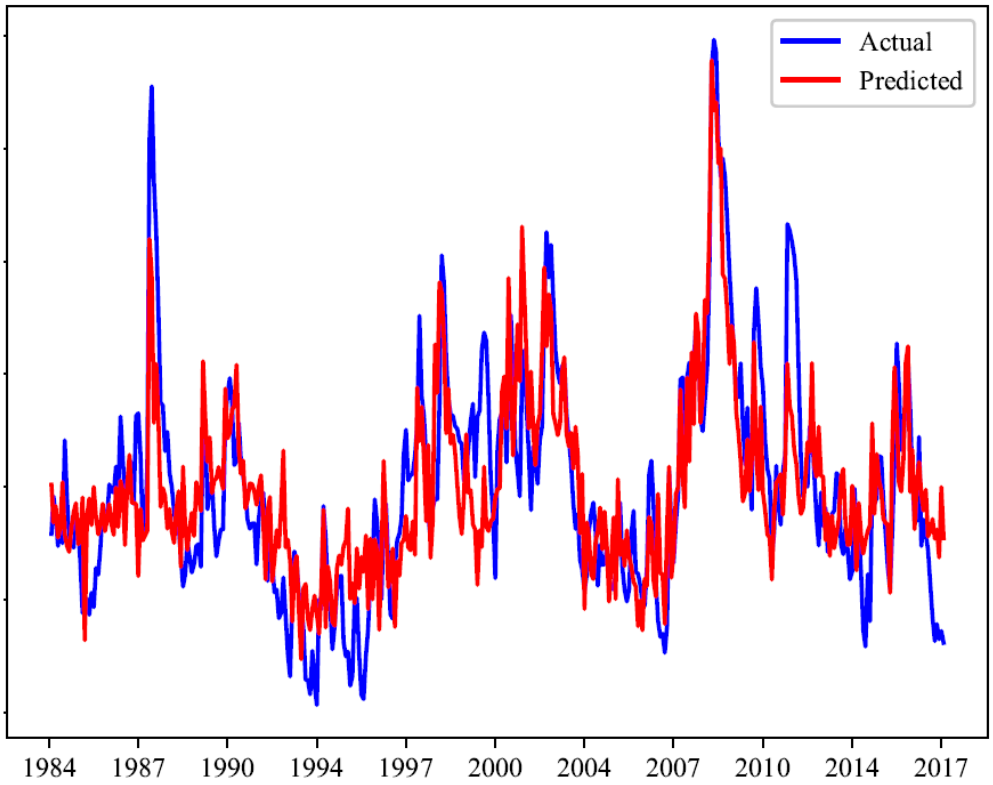


Figure 2.4: Reproduction from the *NBER Working Paper* of Table 2 from Bybee et al. [55], showing the reconstruction of stock market volatility during the period 1984-2017.

sharply to about 65%. Moreover, they found that their measure can explain firms' market-adjusted returns reactions around exogenous weather shocks. Using a sample of extreme weather events (for example hurricanes) from 2003 to 2018 (and controlling for firms' size and book to market values), they found that firms head-quartered in the storm-affected states experienced significant market-adjusted returns of about negative

Figure 2.5: Reproduction from `https://www.policyuncertainty.com/index.html` of the Global Economic Policy Uncertainty (or GEPU) index, during the period 1996Q4-2020Q1.

40 basis points in the $[-5, +5$ day$]$ window around a storm's start date. Notably, firms with higher values of their weather measure was significantly stronger, reaching 130 basis points. Given these outcomes, one should understand why the World Economic Forum's 2018 risk report highlights climate change as one of the most impactful and likely issue facing the global economy[6] and the Financial Stability Board asserts that climate is "the most significant" risk investors will face in the future[7].

Regarding the latter, a more systematic analysis is needed according to the important results reached during time. Loughran and McDonald [207] (hereafter LM) use the *idf* weighting approach, introduced in equation 1.4, in order to construct their DTM (see section 1.5.2) with the aim to measure the tone from $50,115$ firm-year 10-Ks[8] during the period 1994-2008. More importantly, they introduce the idea that when the researcher wants to apply a Sentiment Analysis on a certain Corpus (see section 1.5.1), a suitable *lexicon* should be constructed for that set of documents. In fact, they show that when a standard lexicon is applied to their sample of companies' reports, such as the Harvard-IV-4 TagNeg dictionary, it would classify almost three-fourths of the words as negative, when instead the *same* words are not typically negative in a financial context. Hence, they created a separate list of $2,337$ words with the following attributes: {*negative*, *positive*, *uncertainty*, *litigious*, *strong modal*, *weak modal*}. The paper finds evidence that some word lists are related to market reactions around the 10-Ks filing date, trading volume, unexpected earnings, and subsequent stock return volatility. The LM's dictionary have revolutionized the way in which sentiment time series are calculated in a financial context, and its application range from predicting financial distress in US banks [116], to the analysis of central

[6] `https://www.weforum.org/reports/the-global-risks-report-2018`

[7] `https://www.fsb.org/2018/09/task-force-on-climate-related-financial\protect\discretionary{\char\hyphenchar\font}{}{}disclosures-status-report/`

[8] A 10-K is a comprehensive report filed annually by a publicly-traded company about its financial performance and is required by the U.S. Securities and Exchange Commission (SEC). The report contains much more detail than a company's annual report, which is sent to its shareholders before an annual meeting to elect company directors.

bank communication [288], as we will see in the next section. Anyway, although the *idf* approach is quite intuitive, it does not have to be relevant for market activity. Following this idea, Jegadeesh and Wu (JW) [168] propose a different approach that gives more importance to words that occur on large market move days. In particular, to measure tone, JW create a "global lexicon" by merging multiple dictionaries from the Harvard-IV-4 TagNeg, the Lasswell Value Dictionary [195], the LM dictionary and the word list in [48]. Among other results, they found that a measure of document tone based on this term weighting scheme for 10-Ks is significantly related to filing date returns after controlling for additional factors, such as: (i) earnings announcement date returns; (ii) accruals; and (iii) volatility.

With respect to *Equity research reports*, also here more advanced techniques are used. For instance, Olof [248] uses a convolutional neural network to classify the sentiment in 84,000 research reports written by equity analysts from the equity research department at Skandinaviska Enskilda Banken. He finds that the regime upgrades and downgrades of recommendation are an important feature in order to classify the documents into positive and negative classes. In particular, *logistic regression* (introduced in section 1.2.1) yielded a testing accuracy of 77.84% and a double input channel convolutional neural network yielded a testing accuracy of 83.60%.

Another strand of research focused specifically on the automatic text extraction of company reports features [183] and, more in general, on the *readability* on business documents. Readability is a metric of how easy it is to comprehend text [162]. Given a goal of efficient markets, regulators want to foster transparency by making sure that financial documents disseminated to the investing public are readable. During time, different indices to measure the readability have been provided, such as the Fog index [129], the Flesch-Kincaid Grade level index [304] and the Coleman-Liau Index [74]. The readability of business documents has caught the attention of many researchers, and in particular in the *Accounting* and *Audit* area. Here results seem to be in contrast, in particular with respect to the Fog index: LM found that it may not work well for financial text [205], whilst De Franco et al. [88] combine this index together with the Felsh-Kincaid score, showing that higher readability of analysts' reports is related with an higher trading volume, implying that a better information environment could induce people to trade more and not shy away from the market.

To conclude, regarding the analysis of company filings more in general, it is important to notice that the new trends in financial accountability are promising for future research. In particular, an increasing number of firms worldwide prepare and report their financial statements using the Extention Business Reporting Language (XBLR). The basic idea underlying this language is that it provides a "tag" for every individual item in a company's financial statements, including the notes, which describes the main characteristics of the item. As Palepu et al. suggest in the book "Business Analysis and Valuation" [250], "[...] *by using the* ***appropriate software*** *that recognizes the tags, an analyst can then extract only the needed information from the instance document and ignore irrelevant items. One advantage of XBRL reporting is therefore that it substantially reduces the time that the analyst needs to collect and summarize financial statement information*". We then understand that mining all the XBLR filings with different techniques described in this chapter and in section 1.4 has immense potential, as some work already suggest [151], [200], [256].

2.2 Text Mining and Central Banks' Communication

Despite the increasing interest in text mining in the economic and financial field, it seems that this subject has been historically less used as a technique in the case of monetary policy, macro prudential policy and central banks more in general [37].

This may be due to the large datasets of *quantitative* data that central banks utilize for daily operations, such as the ECB Statistical Data Warehouse[9] (or SDW) or the Federal Reserve Economic Data[10] (or FRED). However, the opportunity to transform texts into quantitative data may be viewed as outweighing the expected benefits [37]. This is even more true in last decades since the way in which central banks operate in the market moved from "secretiveness" to full "trasparency" [282]. The latter feature implies that nowadays a large part of the information on announcement days comes in *verbal* form (rather than quantitative releases) and provides insights to the public by explaining policy decisions, economic outlook, and by shaping market expectations. In this regard, central banks' *communication* has become a key instrument in the central bankers' toolbox [261], particularly during high uncertainty period [73] or when interest rate reaches the zero lower bound [45], [298]. Several channels are usually adopted for such communication, including but not limited to: (i) monetary policy reports; (ii) minutes of monetary policy meetings; (iii) post-meeting briefings by central bank governors; and (iv) speeches by Monetary Policy Committee (MPC) members [321].

Hence, it is important to develop and apply appropriate (text mining) tools to analyse central banks' communications [37]. Following this line, it seems that in last years central banks are "inviting" academics to produce papers or collaborate with them in order to implement new means to improve the quality of monetary policy and financial stability. A practical example of this is the dataset (created from the ECB and available at `https://www.ecb.europa.eu/press/key/html/index.en.html`) for researchers who want to study the content of ECB speech.

According to Picault and Renault [261], we can identify three strands of the literature on central banks' communication, that tries to study: (i) its quantification through textual analysis; (ii) its influence on the predictability of monetary (and related) policies; and (iii) its impact on asset prices and market volatility. We will analyse more in detail all these topics in the following three subsections. Notably, the logic that drives this kind of taxonomy is characterized, in general, by a two part process involving *any* research about central bank communication: the first step is to transform qualitative textual content into quantitative sentiment [76] and/or topics data [182]. The second step is to compare extracted information with respect to traditional (and benchmark) models, such as Taylor rule [315], or those relative to Asset Pricing or Asset Volatility literature. The last goal will be to analyse if central bank communication can improve our understanding of monetary (and related) policy decisions or patterns that we see in financial markets.

2.2.1 Quantifying Central banks' communication

As we stated in section 1.5, the first step of any quantitative analysis is to transform qualitative textual content (in this case for central bank communication) into quantit-

[9] `https://sdw.ecb.europa.eu/`

[10] `https://fred.stlouisfed.org/`

ative structured data. To this purpose, three main methods have been used in the literature [182]: (i) an indirect approach; (ii) manual coding; and (iii) automated textual analysis. In general, these three methods are the basis step to measure both the *direct* effects of central bank communication on economic variables [224], and how it changed (in tone or readability) over time [182], [239].

1. The indirect approach

The indirect approach does not quantify verbal information. Instead, it measures financial market movements in a specific narrow window of decision announcement using high-frequency data [54]. A stylized fact in in these types of analysis is that market's reaction to central bank communication is more pronounced than the reaction to monetary policy decisions. In particular, Gürkaynak et al. [131] investigate the effects of U.S. monetary policy on asset prices using a high-frequency event-study analysis, finding that both monetary policy actions and statements have important but differing effects on asset prices, with the latter having a much greater impact on longer-term Treasury yields. Regarding the ECB, Brand et al. [51] use intra-day changes in money market rates to construct indicators of news about monetary policy arising from policy decisions and official communication, showing that communication may both lead to substantial revisions in expectations of monetary policy and, at the same time, exert a significant impact on interest rates at longer maturities.

2. The manual approach

In recent times, a step further has been made with respect to the indirect analysis in order to identify pieces of information that move the markets. To extract contents, one can follow a manual or an automated approach.

The manual (or human) approach involves the classification of verbal expressions to predefined categories in the same vein of a general text classification task, as described in section 1.4.3. The categories of interest allow the analyst to create quantitative indicators about different topics, such as central banks' opinions about *inflation* [47], [101] or *exchange rate valuation* [95]. Anyway, the most common classification task consists in labelling communication related to *monetary policy* inclinations.

Regarding the ECB, different papers address this kind of analysis. Rosa and Verga [278] manually classified each (ECB's President) introductory statement, according to the tone of the communication, by means of a glossary that translates the qualitative information of the ECB President's into a wording indicator variable. In particular, this indicator takes five values depending on the "*dovish*" or "hawkish" tone of the statement[11]: -2 (very dovish), -1, 0, $+1$, $+2$ (very hawkish). In a similar vein, Picault and Renault [261] hand-coded all sentences in all ECB press conferences between January 2006 and December 2014 in order to build a field-specific lexicon to measure the stance of the ECB monetary policy. They find that quantifying ECB communication using a field-specific weighted lexicon do help in predicting future ECB monetary decision *and* market volatility.

However, the simplicity given from human classification comes at a price with a number of issues. First, the scoring of documents is by definition *subjective* and depends on the analyst. An example of this can be seen in the aforesaid paper from Rosa and Verga [278], where the two researchers disagree with respect to the

[11]The terms Hawkish and Dovish refer to whether central banks are more likely to tighten (hawkish) or accommodate (dovish) their monetary policy.

classification of 14 (over 62) ECB press conferences between 1999 and 2004. Second, (and this could be even more critical) this approach implies a low reproducibility of results. In particular, if classified data are not publicly available, outcomes are not easily reproducible, thus limiting further research and comparability.

3. The automated approach

In order to solve (at least in part) the drawbacks just listed about manual classification, a strand of literature relies on dictionary-based and word-count approaches to ensure that the analysis is transparent and scalable [182].

Regarding the latter, the simplest example can be found in Jansen and De Haan [167], who quantify communication regarding risks to price stability by simply counting the frequency of the word "*vigilance*" in ECB communications. They find that the relationship between the ECB's signalling of inflation risks (as measured from the aforesaid word) and Euro area break-even inflation varied over time. However, the economic significance of this type of communication has been small. Yet, Moretti and Pestre [239] monitored which terms are used in World Bank Reports in the period $1950-2010$, discovering that a major metamorphosis in lexicon has taken place during time, as we can see in Fig.2.6. As the same researchers state, the picture seems to be clear: "*Work in agriculture and industry has been replaced by an overwhelming predominance of financial activities*".

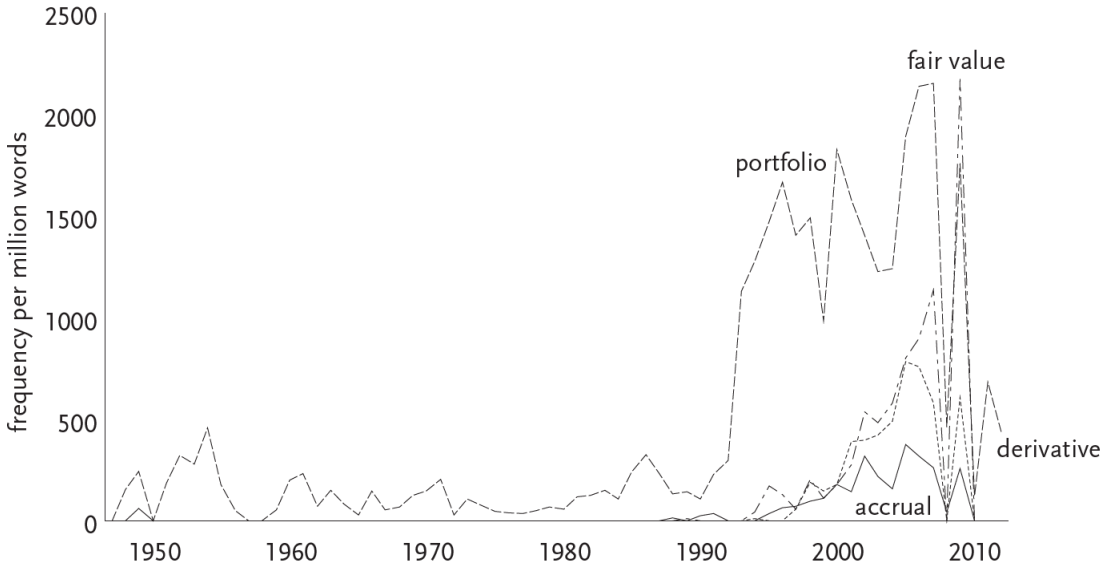


Figure 2.6: Reproduction from the *Stanford Literary Lab* of Exhibit 1 from Moretti et and Pestre [239], showing the rise of a financial language during the period 1950-2010.

In the former case, instead, as we said in section 2.1.3, a researcher predefines a lexicon describing topics of interest. Regarding central banks' operations, several word-lists that separate documents into multiple categories have been created, in particular with respect to: (i) communication measures about hawkish and dovish monetary policy inclinations [61]; (ii) positive and negative tone [136], [10]; (iii) uncertainty [58].

Anyway, it is worth stressing that one of the main difficulties with the dictionary approach is to develop a word-list that *accurately* captures the meaning for a specific application. A not appropriate dictionary may fail to capture all the dimensions and subtlety of central bank communication. This called for the designation of methods that are peculiar to *this* kind of communication, and during years researchers created

suitable dictionaries to apply for specific purposes. Regarding monetary and economic policy, the aforesaid paper from Picault and Renault [261] develops a field-specific dictionary to measure both these two types of stances. It is worth noting that their monetary lexicon allows the analyst to predict future ECB monetary policy decisions in a better way than the one created (for a similar purpose) by Apel and Grimaldi [10]. In Fig.2.7 we reproduce their indicators related to the Monetary Policy and Economic Outlook (last update 06/06/2019) calculated from the ECB press conferences introductory statement since 2006. As we can see, both the monetary and economic outlook dropped after the 2010, one year before sovereign debt crisis in the Euro area. Following a similar logic, Correa et al. [76] construct a dictionary tailored specifically to a *financial stability* context, by means of financial stability reports coming from a panel of 35 countries. The resulting index is showed in Fig. 2.8. Higher values of the index correspond to a deterioration in sentiment: accordingly, one of the spike in the time series coincide to the second Greek Bailout.

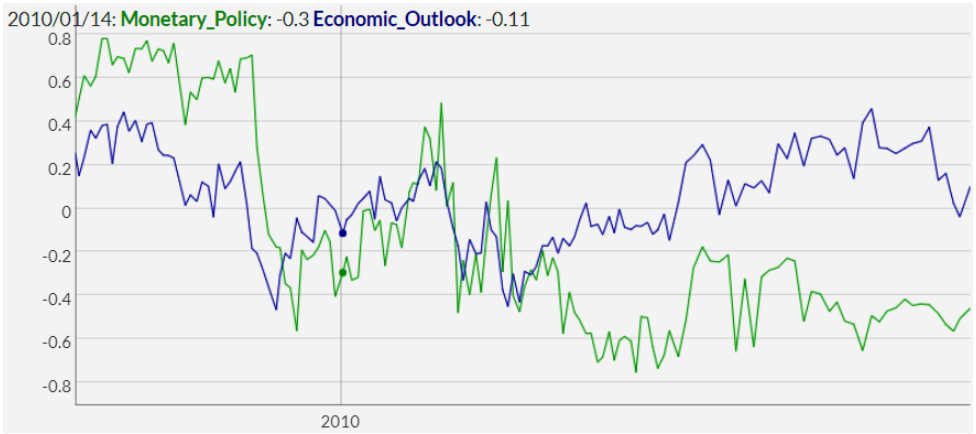


Figure 2.7: Reproduction from `http://cbcomindex.com/index.php` of the Monetary Policy and Economic Outlook Indicator, during the period 2006-2019.

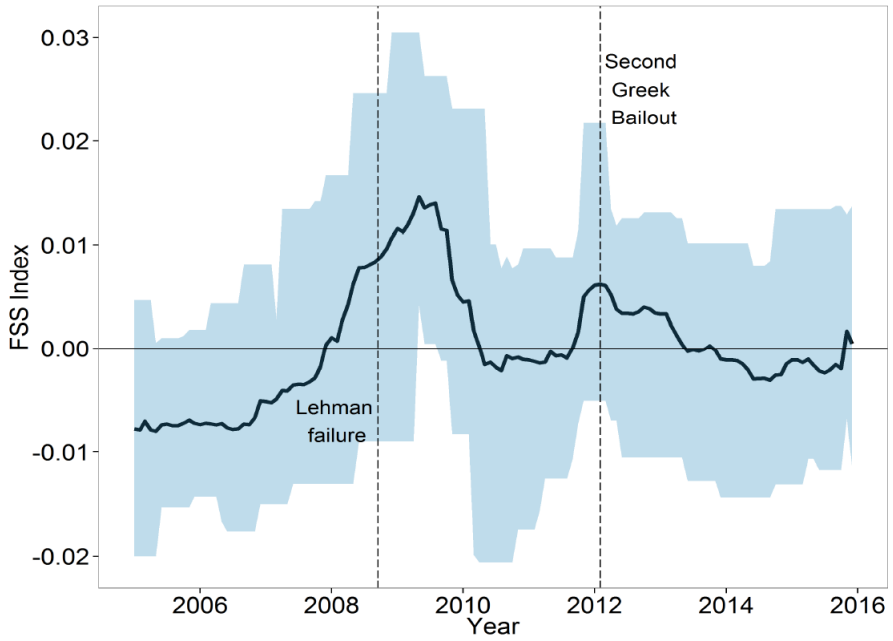


Figure 2.8: Reproduction from the *International Finance Discussion Papers* of Exhibit 1 from Correa et al. [76], showing the Financial Stability Sentiment (FSS) Index during the period 2005-2015.

In this thesis, I will use the two aforesaid dictionaries in order to construct the time series of interest. Specifically, the LM [207] dictionary (see section 2.1.3), together with the Picault and Renault [261] and Correa et al. [76] lexicons will be used to quantifying ECB communication arising from 123 press conferences. I will also construct another dictionary that comes from the union of the Correa et al. [76] and LM [207] lexicons.

2.2.2 Textual analysis of monetary and related policy decisions

All around the world, the objective of almost all central banks is *price stabilization* [222]. Other objectives, that some banks set on the same level of price stability, are related to the *real* ones (such as GDP, occupation, growth, etc.) and, especially after the financial crisis of 2008, *financial stability* [76].

The literature followed closely the expansion over time in the objectives of central banks, and as they rely more on communication to achieve policy objectives, an understanding of their effects is increasingly relevant. In particular, different works test the consistency between "words and deeds" [340] or try to forecast their possible directions, in particular in the post-crisis era [32]. Notably, in most cases it is found that communication conveys information not included in the (quantitative) macroeconomic data.

The strand of literature related to the predictability of future monetary policy decisions is flourishing and involves the analysis of different central banks all around the world. Rosa and Verga [278] (and in a similar fashion Heinemann and Ullrich [144]) show that ECB communication includes information helpful to improve the explanation of interest rate decisions based on a Taylor rule model. A similar result was introduced before in the work of Picault and Renaul [261]. Regarding the FED, Lucca and Trebbi [211] find that short-term nominal Treasury yields respond to changes in policy rates around policy announcements, whereas longer-dated Treasuries mainly react to changes in policy communication. Park et al. [254] quantify the Monetary Policy Board minutes of the Bank of Korea (BOK) using a field-specific Korean dictionary. They find that their lexicon-based indicators help in explaining the current and future BOK monetary policy decisions when considering an augmented Taylor rule.

Text mining is also applied to the area of financial stability. Based on a dataset covering more than $1,000$ releases of Financial Stability Reports and speeches by 37 central banks over the past 14 years, Born et al. [46] find that optimistic FSRs lead to significant and potentially long-lasting positive abnormal stock market returns, whereas no such effect is found for pessimistic FSR. Nopp and Hanbury [247] analyse more than 500 CEO letters and outlook sections extracted from annual reports of 27 banks under the ECB supervision. The analysis of aggregated figures revealed strong and significant correlations between uncertainty in textual disclosures and the quantitative risk indicator's future evolution. In particular, they find that sentiment scores of the textual data reflect major economic events as well as the aggregate Tier 1 capital ratio evolution. Bholat et al. [36] analyse confidential letters sent by the Bank of England's Prudential Regulation Authority to banks and building societies it supervises using a supervised ML method (see section 1.2.1) . They find that, in terms of negative words and direct languages, the letters vary according to the riskiness of the firm. In conclusion, the aforesaid work from Correa et al. [76] shows that their FSS index deteriorates just prior to the start of banking crises. In this Thesis, the

financial stability ECB tone will be analysed using the Correa et al. [76] dictionary.

Lastly, several works focus on the effects of central bank communication on the real economy. Analysing the tones of 22 central banks during $2000 - 2015$, Luangaram et al. [210] find that the Federal Open Market Committee's communication tone on the real economy was consistent with their interest rate decisions, and broadly in line with actual GDP growth. The tone of the ECB's Governing Council, on the other hand, was predictive of both growth and inflation outlook. Hansen and McMahon [136] explored the channels through which FED communication has effects finding that FED guidance on future interest rates seems to have been more importance than their communication about economic conditions. Kawamura et al. [178] conducted a discourse analysis of the Bank of Japan's Monthly Report and examined its characteristics in relation to business cycles. They found that difference between the number of positive and negative expressions leads the leading index of the economy by approximately three months.

2.2.3 Central Banks' communications and Financial Markets

Central bank (and more in general monetary policy) announcements are among the most impactful events to global financial markets [261], [288]. In particular, while real macroeconomic variables (such as GDP and employment) respond to the effects of policy innovations over a long horizon, the financial markets anticipate future changes to the real economy and financial asset prices react *instantaneously* [169], as we saw early in Fig.2.1b.

The pioneer work in this area comes from Sadique et al. [283] which show that financial markets do respond to FED announcements. In particular, they examine the relation between S&P 500 Index returns and Beige Book tone. They do find a significant relation between Beige Book tone, volatility and trading volume. These results are confirmed from the work of Smales and Apergis [301], that show also that markets are more responsive to monetary policy language and decisions during recessions. In a more recent paper, Brusa et al. [54] show that "Fed exerts a unique impact on global equities". Specifically, studying how and why announcement risk premia vary globally across Federal Reserve, ECB, Bank of England and Bank of Japan during the 1998–2016 period, they find that high equity returns around monetary policy announcements are a phenomenon that is unique to the FED. As the same research notice, for the non-US central banks, there is no announcement premium even in their home market, as we can see in Fig. 2.9.

Regarding the ECB, Schmeling and Wagner [288] use the LM dictionary to analyse 209 press conferences during the period 1999-2017. They found that, during press conferences days, a more positive (negative) tone is associated with: (i) higher (lower) equity market returns (as proxied from the Euro Stoxx 50); (ii) lower (higher) volatility risk premia; and (iii) lower (higher) credit spreads (and viceversa). Regarding the first result, in chapter 4 I will show that taking into account the LM dictionary with a financial stability lexicon [76] can improve our understanding of how ECB communication impacts on the European stock market. Klejdysz et al. [182] study how the dynamics of topical composition of the ECB press conference affects stock market volatility on the Governing Council meeting days. They find that market uncertainty, as proxied from the Volatility Derivatives on Eurex Exchange (or VSTOXX) index, increases if the ECB switches to a different communication topical regime, a pattern that we clearly see in Fig.2.10.

Similar results are empirically shown for other banks all around the world. Mathur

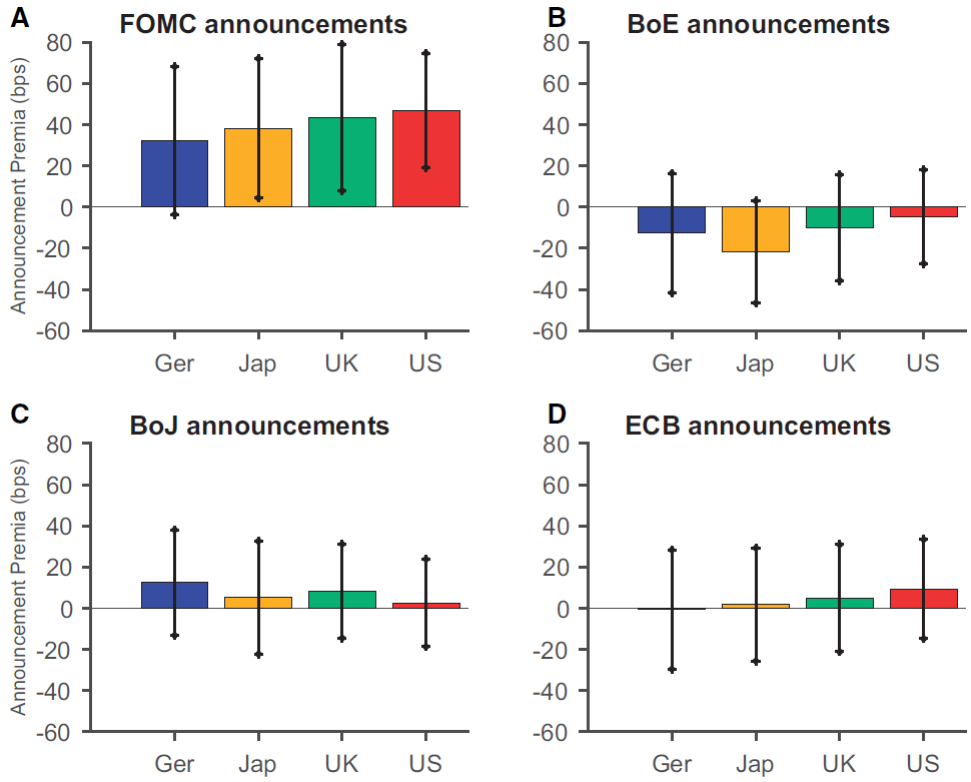


Figure 2.9: Reproduction from the *Review of Finance* of Exhibit 1 from Brusa et al. [54], showing the average stock market excess returns for the countries associated with our four major central banks over a 2-day window, during the 1998–2016 period.

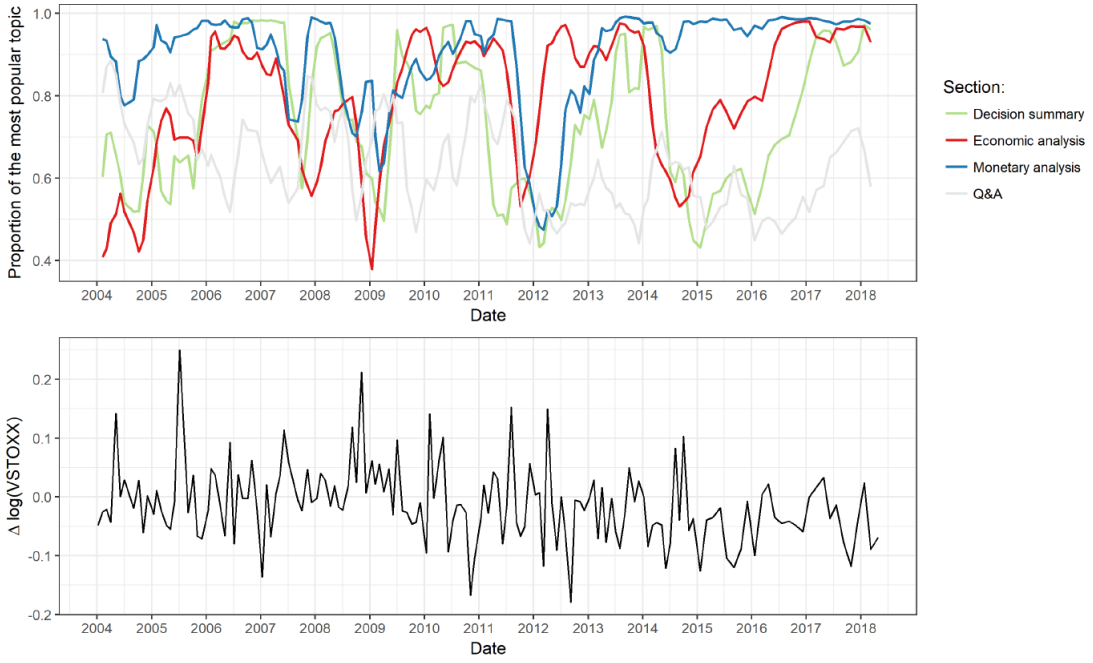


Figure 2.10: Reproduction of Exhibit 5 from Klejdysz et al. [182], showing the daily percentage change (close to close) of the VSTOXX on the day of the ECB press conference.

and Sengupta [224] quantitatively analyse monetary policy statements of the Reserve Bank of India from 1998 to 2017. Using an ordinary least squares regression, they find that lengthier and less readable statements are linked to both higher trading volumes and higher returns volatility in the equity markets, though these effects are

not persistent over time. Other contributions include those by Ranaldo and Rossi [271], who use intra-day asset price data and find significant price effects of Swiss National Bank communication on bond, currency, and equity markets and the one from Hendry and Madeley [146], who investigate the kind of information from Bank of Canada's communication statements over the 2002-2008 period by means of a Latent Semantic Analysis methodology. They find that discussions about geopolitical risk and other external shocks affect market returns and volatility, especially for short-term markets, such as the one related to the Three-Month Canadian Bankers' Acceptance Futures.

Chapter 3

Central Banks Policies and Herding Behaviour in Financial Markets

According to Hirshleifer and Hong Teoh [148], individuals' (and in particular investors') thoughts, feelings and actions can be influenced by other individuals in several ways: (i) by observation of *actions* (e.g., observation of quantities such as supplies and demands), (ii) by observation of the *consequences of actions* (such as individual payoffs, or market prices) and (iii) by ***words***.

This fact is even more true if we consider central banks since, as we highlighted in section 2.2, *communication* became a key instrument in their toolbox [261], particularly during high uncertainty period [73] or when interest rate reaches the zero lower bound [45], [298]. An empirical case of how investors seem to trade in a similar manner in response of an ECB's President statement (that this, to *herd*) was shown in Fig. 2.1. There, we can clearly see how the trading volume (considered to be a good proxy to measure investor herding [33]) of that day was quite higher with respect to other days in which the European stock market showed more normal values.

At the same time, it also true that this kind of investors' behaviour is not only affected from central bank communication: looking at Fig. 2.1a, we can notice how a Facebook's CEO *sentence* created grasp in financial markets, and, again, a similar trade behaviour (as proxied by the abnormal trading volume of that day).

Hence, the aim of this chapter is to shed light on the *rational* and *irrational* reasons of herding behaviour in financial markets. Specifically, in section 3.1 we will frame this concept in order to understand its *sources*, *metrics* and *impacts* in a financial context, with particular emphasis on the *equity* market (also if other capital markets would be relevant, like for instance, the *insurance* one [204]). Then, in section 3.2, we will look at how central bank can influence (and be influenced from) such a type of pattern. In particular, we will analyse how policy makers can apply a regulation that takes into account behavioural insights in order to prevent investor herding on an *ex-ante* basis [16]. This regulatory framework is of particularly relevance for central banks, since, as we will show in this chapter, it could be one of the main drivers behind the instability of the whole financial system.

3.1 Herding and financial markets

Generally speaking, in an economic and financial context, with the term *herding* (or herd) behaviour we mean the process where economic agents are imitating each other actions and/or base their decision upon the actions of others [307]. In other words, demand fuels further demand, or supply fuels further supply [176].

The academic research toward the study of this type of behaviour (that seems to be initiated from the 1990s onwards) stemmed, above all, as a response to the accelerated process of globalization, that implied (mainly financial) links between markets at the international level. In fact, as Forbes suggests in his book "Behavioural Finance" [113]: "*the fact that we observe herding in financial markets should not really surprise us, given the extent of conformity in both product markets and investment or the productive technology to make good in high demand*".

Although some economists argued that herding may results in efficient outcomes [294], it is widely accepted the opinion that considers this kind of behaviour as one of the main driver behind several (and problematic) patterns frequently observed in financial markets, such as periods of extreme volatility [41] or bubble-like episodes [30]. Regarding the latter, in particular, herding as a behavioural trait of investors has been persistent identified throughout centuries, with episodes that span from the *Tulip Bubble* in Netherlands in the late 1630s (well documented in the famous book "*Confusión de Confusiones*" [90]) up to more recent events such as the Dot-Com bubble[1], the 2008 global credit crisis [160] and the actual Covid-19 global crisis [5], [26], [346].

Hence, it is then extremely important (and in particular for policy makers, as we will see in the next section) to understand both the *sources*, *metrics* and *impacts* of such type of investor behaviour. Following this line, according to Spyrou [307], it is possible to divide the discussion about herding in financial markets trying to addressing the following three questions: (i) Which are the theoretical explanation behind herding behaviour in financial markets?; (ii) Which are the metrics used in order to quantitative measure herding?; (iii) Which is the empirical evidence of herding in financial markets?. Following this order, the answers to these questions will be investigated in section 3.1.1, 3.1.2 and 3.1.3, respectively. Moreover, in order to integrate the taxonomy provided by Spyrou [307] (in light also of more recent findings) different papers offering a systematic examination of this topic have been analysed: [130], [148], [175], [184], [190], [294], [309], [327].

3.1.1 Explaining the sources of herding in Financial Markets

Explaining why herding behaviour occurs in financial markets is one of the most challenging tasks to address in Economics and Finance. Still nowadays, results seem to be in contrast, and one of the main reasons for that can be attributable to the two different schools of thought that try to theoretically explain the sources of herding. In particular, they can be divided between the ones that see this phenomenon as a *rational* [94] or *irrational* behaviour [148].

It is worth noting that although the distinction between rational and behavioural herding eases conceptualisations, their effects may be intertwined and it is frequently difficult to discern between the two types of herding behaviour observed in real markets. It is then important to consider both [134].

[1] `https://nickledanddimed.com/2018/08/01/herding-and-the-dotcom-bubble/`

1. Rational herding in financial markets

Under certain circumstances, it seems that herding is a rational (and sometimes an obliged) choice, and (as we will see soon) the rationales for this can be the most diverse [94]. It is important to notice that rational herding models are part of the wide range of models in behavioural finance which consider a *representative agent* [138] as the "average" investor operating in financial markets. The underlying assumption of these kind of models is that all investors have the same attitude to risk and face the same reward to alternative courses of action. Investors only differ in their place in queue of decision markers.

Rational herding can be further divided into "*intentional*" and "*spurious*" herding [175], [307].

Intentional herding refers to the situation where investors mimic their peers in order to (try to) exploit a *payoff* (or gain) that they could not realize themselves in the absence of such imitation [40]. The literature proposed two different types of payoffs that can drive such a type of herding.

The first one refers to *informational* payoff, where investors decide themselves to copy the trades of their peers in order to free-ride their private information [70]. This situation is also referred to as "informational cascade" [148]. The rationale behind such a decision is the investor's (awareness of) effective lack of information with respect to her peers or her lack of ability in processing them. Actual examples of such situation are the *copy trading* platforms that allow (mostly retail) investors to copy the strategies of most famous fund managers. The possible effects of such business attracted the attention of several researchers, in particular with respect to the changes in risk appetite that these trading strategies imply [11], [34].

The second one refers to *professional payoff* [287], in particular among fund managers [229] and (equity research) analysts [327]. In this case, as Kallinterakis and Gregoriou [175] suggest, the key point is to notice that investment professionals' performances are valued in *relative* terms (that is, they are assessed versus the performance of their peers). This situation could be the reason for which "younger" equity research analysts know that if they make bold forecast deviating from market consensus they are more likely to be fired [331], or why money managers may mimic the actions of other money managers in order to preserve their *reputation* [155], [213]. This type of herding is particularly relevant during periods of market downturns [214], when the analyst or a manager can blame their losses on the overall adverse state of the market.

However, in both situations we can easily understand that intentional herding is detrimental to the EMH [215] (see section 2.1) since the imitation between investors' trading strategies imply a temporary blockages of information, thus deterring their incorporation into asset prices. Herding thus leads to an inefficient market situation characterized by market fluctuations [130]. Moreover, it has been documented that besides the creation of fragile markets [132], intentional herding could be also the cause of financial markets volatility [41] and systemic risk [16], [318].

Spurious herding, instead, occurs when investors exhibit similar reactions to *commonly* observed signals. According to Sharma and Bikhchandani [294] this type of herding is an efficient outcome and the examples proposed in their paper well explain the definition just given. For instance, if interest rates suddenly rise and thus stocks become a less attractive investments, it is likely that many investors may want to hold a smaller percentage of stocks in their portfolio. This example does not refer to an intentional herding scenario, since in this case investors are not revising their

decisions upon peers' strategies. Instead, they are reacting to *commonly* known public information (i.e. the rise in interest rates).

Also in the case of spurious herding, the literature proposed two possible sources for such a kind of behaviour [175].

The first one relates to *relative homogeneity*, for which the presence of features that are common to a broad group of investors imply a similar behaviour when they have to face a similar (financial) decision problem [316]. Two simple examples of such sources are the similar educational backgrounds of fund managers (that could imply a similar valuation of economic indicator [231]) and the regulatory framework to which they are subject (that could lead to a similar portfolio composition [268]).

The second source refers to the similar investment strategy adopted across professional investors, thus adopting a *similar type of style investing* [28]. Style investing implies that managers will select assets in their portfolio according to particular characteristics, that span from past performance (i.e. momentum [91]) to religious or ethical principles (ex. Environmental, Social and Governance, or ESG, criteria [232]). Hence, if several investors pursue a specific strategy, then it will be likely that their trades will show positive correlation without *any* interactive observation having occurred: any similarity in their trades could be just the result of their common use of the type of the adopted style analysis.

2. Irrational herding in financial markets

Less-than-perfectly rational factors have been found to be relevant to herding, either in conjunction with the aforementioned intentional/spurious ones or in isolation [175], although in this case the strand of research seems to be less flourishing with respect to rational herding.

However, it is worth noting that typical sources implying irrational herding behaviour has been found to relate to *behavioural forces*, such as the *availability heuristic* [290] and the tendency of investors to prefer to select stocks whose companies' headquarters are in close proximity to their country of residence (i.e. home bias [78]). Regarding the latter, in particular, such a tendency has been found to amplify the correlation in trades among home biased investors [291].

Yet, non-rational herd behaviour can arise as the consequence of *psychological stimuli* [19], such as pressure from social circles [18], social conventions [17] and/or cognitive dissonance [265]. For example, Keynes, in his book "*The general theory of employment, interest, and money*" [180] argues that investors are affected by sociological factors (leading to the formation of the so-called *animal spirits* [221]) that may drive market participants to imitate each other during periods of uncertainty. As we said in the introduction of section 3.1, such irrational behaviour may give rise to bubble-like phenomena [295].

3.1.2 Quantitative measures of herding in Financial Markets

In general, we can classify empirical methodologies to measure herding in financial markets into two main categories [307]: (i) studies that rely on micro (or proprietary) data, analysing whether *specific* types of investors herd; (ii) and studies that instead rely on aggregate prices and market activity data, thus investigating herding toward the market consensus.

Now, both these two kinds of approaches will be discussed. However, more attention will be paid to the latter given its usage (of one of the sub-methodology) in Chapter 4

of this Thesis.

1. Empirical methodologies that use private-data

Early work about herding behaviour in financial markets focused on the analysis of *money managers*. In particular, Lakonishok et al. [191] (henceforth LSV) use a dataset based on 769 all-equity pension funds between 1985 and 1989 to evaluate the potential effect of their trading on stock prices. In order to measure herding behaviour, they introduce their herding measure H for a given stock in a given quarter, defined as follow:

$$H(i) = \left| \frac{B(i)}{[B(i) + S(i)] - p(t)} \right| - AF(i). \tag{3.1}$$

Where:

- $B(i)$ is the number of money managers who are net buyers,
- $S(i)$ is the number of money managers who are net sellers,
- $p(t)$ is the expected proportion of money managers buying in that quarter relative to the number active,
- $AF(i)$ is the expected value of $\left| \frac{B(i)}{[B(i)+S(i)]-p(t)} \right|$ of no herding.

The logic behind LSV's herding measure is simple: if there is a tendency of money managers to disproportionately buy (sell) an individual stock, then it can be concluded that there is herding at the level of individual stocks. In particular, LSV point out that, if no herding exists then the value of $AF(i)$ should not vary from period to period (on the other hand, in the presence of herding there should be significant cross-sectional variation in this measure).

In a similar fashion, in order to measure the herding of *institutional investors*, Sias [296] estimates the cross-sectional correlation between the demand for an asset by institutional investors last quarter and demand for the asset by institutional investors the next quarter. Specifically, he calculates at the beginning of each quarter each institutional investor's position in each security as a fraction of the security's shares outstanding. Intuitively, an institutional investor is defined as a buyer (seller) if their ownership increases (decreases) in the next quarter. Hence, in order to estimate the portion of investors that are buyers, he calculates the fraction of institutional investors trading the security k that are buyers in the quarter t as:

$$Raw\Delta_{k,t} = \frac{BI_{k,t}}{BI_{k,t} + SI_{k,t}} \tag{3.2}$$

where $BI_{k,t}$ is the number of institutions buying asset k during quarter t, and $SI_{k,t}$ the number of institutions selling asset k during quarter t.

Thus, to allow for an aggregation over time and comparison for different market capitalizations and investor types, Sias standardizes the above measure as follows:

$$\Delta_{k,t} = \frac{Raw\Delta_{k,t} - \overline{Raw\Delta_{k,t}}}{\sigma(Raw\Delta_{k,t})}$$

In conclusion, he calculates a cross-sectional regression defined as:

$$\Delta_{k,t} = \beta_t \Delta_{k,t} + \varepsilon_{k,t}$$

The intuitive idea behind this regression is that if institutional investors follow each other, then we would expect that the fraction of institutions buying the asset k in the current quarter t is positively correlated with the fraction of institution buying that asset in the previous quarter.

The main difference between the measures introduced in eq. (3.1) and (3.2), as Sias points out, is that while the former indirectly tests for cross-sectional temporal dependence during time, the latter provides a direct test of whether institutional investor herd during the following periods.

2. Empirical methodologies that use market-data

Empirical investigations of herding behaviour in financial markets is even more vibrant, and different metrics have been proposed during time.

The pioneer work in this regard can be found in the measure introduced from Christie and Huang [72] (hereafter CH) which suggest that the investment decision-making process used by market participants depends on overall market conditions (or *sentiment*). In particular, during normal periods, *rational* asset pricing models will predict that the dispersion in cross-sectional return will increase with the absolute value of the market returns, given the fact that in this case investors are trading on their own information. On the other hand, during periods of extreme market movements, investors will tend to discard their own private information, and their investment decisions will be more likely to replicate collective actions in the market: *individual* stock returns will tend to cluster around the overall *market* return. Formally, to measure such a return dispersion, CH introduce the cross-sectional standard deviation (or $CSSD_t$), expressed as:

$$CSSD_t = \sqrt{\frac{\sum_{i=1}^{n}(R_{i,t} - R_{m,t})^2}{N-1}}. \tag{3.3}$$

Where:

- $R_{i,t}$ is the observed stock return of industry i at time t,
- $R_{m,t}$ is the cross-sectional average stock of N returns in the portfolio at time t,
- N is the number of firms in the portfolio.

As CH suggest, lower values for $CSSD_t$ imply a higher herding activity in the financial market.

However, in a later study, as Chang et al. (hereafter CCK) [63] noted, since the $CSSD_t$ tends to be sensitive to outliers, they proposed the cross-sectional absolute deviation (or $CSAD$), defined as follows:

$$CSAD_t = \frac{1}{N}\sum_{i=1}^{n} |R_{i,t} - R_{m,t}| \tag{3.4}$$

Moreover, using this measure, they implement the CH idea by saying that rational asset pricing predict both the fact that equity return dispersions are an increasing function of the market return and that this kind of relationship is *linear*. Nevertheless,

in periods of herding activity, if investor tend to show common behaviour, the relation can become non-linearly increasing or even decreasing [63] in the $CSAD_t$ measure. To test this in a statistical way, they build a regression using the conditional version of the CAPM [38] in order to establish the presence of a linear relation between $CSAD_t$ and $R_{m,t}$. In particular, to allow for the possibility that the degree of herding may be asymmetric in different state of the markets (i.e. up and down periods), they ran two different empirical specifications:

$$CSAD_t = \beta_0 + \beta_1^{UP} \left|R_{m,t}^{UP}\right| + \beta_2^{UP}(R_{m,t}^{UP})^2 + \varepsilon_t \tag{3.5}$$

$$CSAD_t = \beta_0 + \beta_1^{DOWN} \left|R_{m,t}^{DOWN}\right| + \beta_2^{DOWN}(R_{m,t}^{DOWN})^2 + \varepsilon_t \tag{3.6}$$

Subsequently, Chiang and Zeng [66] (CZ) develop eq. (3.4) in order to run a test for detecting herding activity implementing the following regression:

$$CSAD_t = \beta_0 + \beta_1 R_{m,t} + \beta_2 \left|R_{m,t}\right| + \beta_3 R_{m,t}^2 + \varepsilon_t \tag{3.7}$$

Despite the similarity between eq. (3.5 - 3.6) and (3.7), CZ explain that their measure differ in that CCK's measure was grounded (as we said before) on the conditional version of the CAPM. CZ measure, instead, follow the (above mentioned) CH procedure, in order to avoid possible specification error associated with a single-factor capital asset pricing model. Moreover, they show that:

- $\beta_1 + \beta_2$ captures the relation between market and asset dispersion when $R_{m,t} > 0$,
- $\beta_1 - \beta_2$, instead, captures the above relation when $R_{m,t} \leq 0$,
- The ratio $(\beta_1 + \beta_2 / \beta_1 - \beta_2)$ is the relative amount of asymmetry between stock return dispersion and the market's return.

The high degree of reproducibility of the CZ methodology implied, in the following years, a development of similar (and augmented) regression. For instance, Belgacem and Lahiani [31] in order to test the intensity of herding behaviour around US macroeconomic announcements days, augment eq. (3.7) by including eleven US macroeconomic releases as follow:

$$CSAD_t = \beta_0 + \beta_1 R_{m,t} + \beta_2 \left|R_{m,t}\right| + \beta_3 R_{m,t}^2 + \sum_{k=1}^{11} \beta_k D_k R_{m,t}^2 + \varepsilon_t. \tag{3.8}$$

Where D_k is a dummy variable taking the value 1 on the days of k^{th} news announcements, and 0 otherwise. The term $\sum_{k=1}^{11} \beta_k D_k R_{m,t}^2$ allows them to detect herding behaviour around 11 US macroeconomic indicators.

Likewise, in the last chapter of this Thesis, I will augment eq. (3.7) in order to assess the evidence of whether herding behaviour of investors occurs around ECB press releases announcement days. Specifically, I will consider if herding behaviour occurred around negative and positive spikes of my sentiment time series derived from ECB statements. Similarly to my work, Ren and Wu [275] proposed a method to detect herd behaviour by means of a sentiment analysis approach. In particular, sentiment indexes are utilized to substitute market return in eq. (3.5 - 3.6). As they suggest, detecting herding in this way could improve our understanding of such a phenomenon

given the fact that it is highly linked with *human psychology* [82] and *sociology* [255], [267]. Yet, Nobel Prize Winners Akerlof and Shiller [4] assert that "*we will never really understand important economic events unless we confront the fact that their causes are largely mental in nature*". In the same vein, my empirical model will build on both these kinds of thought.

To conclude, it is worth noting that another type of procedure (and quite different with respect to the ones presented so far) is the so-called *beta herding*. This measure has been proposed by Hwang and Salmon [163] (HS) and an important feature is that it can capture investor herding as a time-varying phenomenon. In particular, their approach is grounded on the idea that investors suppress their beliefs regarding equilibrium, which is subsequent reflected on individual stock betas that converge towards the market beta. Specifically, the relationship between the beta in asset pricing equilibrium $\beta_{i,m,t}$ and the biased beta $\beta^b_{i,m,t}$ can be described as:

$$\beta^b_{i,m,t} = \frac{E^b_t(r_{i,t})}{E_t(r_{m,t})} = \beta_{i,m,t} - h_{m,t}(\beta_{i,m,t} - 1). \tag{3.9}$$

Where $r_{i,t}$ and $r_{m,t}$ are the excess stock return of asset i and the the market return at time t, respectively, and $E_t(.)$ is the expectation operator conditional on the information set at time t.

Hence, the mispricing attributable to cross-sectional bias is reflected in the superscript b, and the level of herding in betas due to the mispricing in $h_{m,t}$. It then follows that:

- If $h_{m,t} = 0$, prices are in equilibrium and there is no herd behaviour,
- If $0 < h_{m,t}$, we have adverse herding,
- If $0 < h_{m,t} < 1$, there is evidence of herd behaviour towards the consensus
- If $h_{m,t} = 1$, perfect herd behaviour is suggested, and thus asset prices are expected to move towards the consensus (i.e., the market portfolio).

As for the CCK's measure [63], also the HS must be implemented on the choice of an appropriate asset pricing model.

3.1.3 Empirical evidence of herding in Financial Markets

During years, developments in herding studies increased (in particular as research area of behavioural finance [175]), and, together with them, the focus narrowed on specific subjects and/or patterns. In particular, we can classify the related work that try to explain herding behaviour across (i) institutional investor, (ii) analysts' recommendations, (iii) aggregate markets activity and (iv) specific sectors.

1. Institutional investor herding

Early work analysing herding behaviour of institutional investors has grown since the Lakonishok et al. [191] work cited in section 3.1.1. Applying the procedure introduced in eq. (3.1), they found weak evidence of herding, and somewhat stronger evidence of positive-feedback trading in smaller stocks (however with no destabilizing effect on stock prices). In particular, they conclude that there is no solid evidence in the data that institutional investors destabilize prices of individual stocks.

Some years later other studies attempted to study the same phenomenon. The first refers to the one of Grinblatt et al. [127], that analyses, by means of the LSV measure, quarterly portfolio holdings for 274 mutual funds during the period 1974-1984. They found evidence of momentum strategy across investors, but they also concluded that the relation between a fund's tendency to go with the herd and its performance largely disappeared after controlling for the fund's tendency to buy past winners. Another one, instead, refers to the work of Nofsinger and Sias [246], which in a opposite way from Grinblatt et al. [127], document strong positive correlation between changes in institutional ownership and returns measured over the same period in the US market. In particular, they found that herding behaviour could be explained both from the fact that institutional investors positive-feedback trade more than individual investors and that institutional herding impacts prices (as one would expect) more than herding by individual investors. The third study is from Sias [296], that found evidence of herding in the US market by means of his measure introduced in eq. (3.2). In particular, by observing that the fraction of institutions buying in a certain quarter positively co-varied with the fraction of institutions buying in the next quarter, he advanced the hypothesis that this phenomenon could be attributable to the fact that institutional investor might infer information from each other's (that is, they could be affect from *intentional herding*, introduced in section 3.1.1).

In more recent years, another strand of research focused only on the analysis of herding activity of *hedge funds* managers and their related effects on financial markets. For instance, Haigh et al. [132] utilize a dataset from the U.S. market on individual positions of speculative traders in 32 futures markets covering the period of time 2002 - 2006, in order to understand if hedge funds herding could be the cause behind price destabilization in those markets. However, despite the fact that they found some evidence of herding amongst hedge funds and other types of speculators, they concluded that hedge funds herding has not been destabilizing [171]. In a similar study, Jiao and Ye [171] analyse whether hedge funds and mutual funds tend to herd together and if they show this kind of behaviour after an important investor transaction. They found evidence of mutual funds herding into or out of stocks following the herd of hedge funds, but not the opposite. Moreover, as Haigh et al. [132], they also show that hedge fund herding itself does not destabilize prices.

2. Herding across analysts' recommendations

As we said earlier in section 3.1.1, several theories suggest that career concerns [15] (e.g. reputation, compensation) or the imitation of analysts with higher capabilities [154] could imply an under-weighting of private information and the subsequent herding behaviour toward the consensus. Analyst herding is of particular interest for financial markets, since their forecasts are often used as one (perhaps the main) information source by other investors. Therefore, deviations from optimal forecasts could be troublesome for the efficiency of the financial market [327].

These theories seem to find support thanks to the results of empirical studies. For instance, Hong et al. [154] use the Institutional Brokers Estimate System (or I/B/E/S) to get data of $8,421$ US security analysts who produced earnings forecasts between 1983 and 1996. In order to examine how their forecasting ability is related to career concerns, they found that inexperienced analysts are more likely to be terminated for bold forecasts that deviate from the consensus. Moreover, they also found that younger analysts tend to herd more compared to their experienced counterparts. However, this does not seem to be a general result: Ashiya and Doi [14] find that macroeconomic

forecasters in Japan seem to herd regardless of the age of the forecaster.

On the other hand, Bernhardt et al. [35] argue that the clustering of analysts' forecasts and recommendations may not necessarily imply analyst herding. For example, Zitzewitz [353] utilize the I/B/E/S finding that analysts tend to exaggerate (or anti-herd) their differences with the consensus. This result is robust to different specification and it is present in nearly all sub samples of the data. Similarly, Pierdzioch et al. [262] analyse 20,000 forecasts of nine metal prices at four different forecast horizons. They found anti-herding strategy appears to be a source of forecasts' heterogeneity. As they suggest, forecasters anti-herding reflects strategic interactions between them in order to foster incentives and scatter forecasts around a similar consensus.

3. Herding and aggregate markets activity

A strand of the literature attempts to uncover herding behaviour using aggregate market data, in particular after the different methodologies illustrated earlier. For instance, Christie and Huang [72] use daily data for NYSE and Amex firms during the period from July 1962 to December 1988. However, their results show that daily and monthly returns are inconsistent with the presence of herding during periods of large price movements.

Chang et al. [63] expand the analysis at international level, analysing whether herding behaviour was observed in the US, Hong Kong, Japan, South Korea, and Taiwan equity markets. They found no evidence of herding in the US and Hong Kong in Japan market. Instead, results showed that emerging markets (there represented from the South Korea and Taiwan) showed patterns of herding behaviour. It is worth noting that this finding is not confuted from subsequent empirical research and that herding seems to be a phenomenon more intrinsic to emerging markets than to developed markets [175]. Specifically, it has been found that investors seem to herd more in markets such as Egypt [230], Iran [238], India [193], [216], [266], Kenya [64], Marocco [143], Mongolia [103], Pakistan [170], Poland [329], Portugal [152], Russia [164], South Corea [69], Taiwan [156], Tunisian [135], rather than in the European Union [310], UK [342], and US [72] stock market.

As Gelos and Wei [117] pointed out, this spread evidence should be ascribed to the relatively lower transparency of emerging markets which (in turn) renders the quality of public information questionable, thus prompting institutional investors to mimic each other when trading there.

Other authors link herd behaviour with the creation of asset bubbles. In particular, as Dass [86] suggest, investing in popular or "hot" stocks during periods of high market volatility indicates investing in identical assets as other investors, implying herd behaviour. Moreover, Johansen and Sornette [174] studying the Japanese stock market argue that herd behaviour by investors not only leads to speculative bubbles, but also the so-called "anti-bubbles". Anti-bubbles are present when market valuations fall after the all-time high levels achieved during the bubble (i.e. a bearish phase).

In conclusion, it is important to notice that stock capitalization is another important driver for herding behaviour [296], [161]. Small capitalization stocks have reported stronger herding behaviour, and this could be ascribable to their higher informational risk, in particular in the banking sector [3], [63].

4. Herding on specific industries or sectors

The possibility of both managers and retail investors herding at the industry or sector level has been explored by a series of studies. The work in this field is justified from the fact that investors usually base their trade decisions on sector-specific information. For instance, money managers often make portfolio recommendations only at the sector level. Therefore, sector-specific market data form a natural ground for testing of herding behaviour [71].

Early study comes from the work of Choi and Sias [71]. They apply the LSV measure defined in eq. (3.2) in order to test whether institutional investor in US herd among 49 industries, as identified from Fama and Frech [104]. They found that institutional investors did herd into and out of the same industries, justifying this results in line with reputation preservation (see section 3.1.1) and style investing hypothesis [28]. Also Litimi et al. [202] analyse the US stock market at a sectoral level. Granger causality test shows that herding is an important driver behind the possibility of bubbles in some sectors, but not all. In particular, the sectors is which herd in detected are the Industrial, Health&Care, and Public utilities.

However, sectoral herding is a phenomenon present also in other countries, both in emerging [56] and developed [249] ones. In particular, Cakan and Scranton [56] apply the methodology introduced in eq. (3.5 - 3.6) to the Turkish sectoral daily stock prices from 2002 to 2014. Notably, they detect herding behaviour in the financial and technology sector during highly volatile markets. Instead, low-volatility markets show herding behaviour only in the service sector. In a study that focus on the European equity market, Ouarda et al. [249] analyse 174 shares (with monthly frequency) from the Euro Stoxx 600. They found that herding is present across the majority of sectors (i.e. Oil&Gas, Basic Materials, Financials, Industrials, Health&Care, Consumer services, Telecommunication, Utilities and Technology) but the same cannot be said for the Consumer Good sector.

Similarly to Ouarda et al. [249], in this Thesis, I will study whether herding behaviour occurred in six different sectors within the European equity market. However, my study differs from them since I focus on the 50 component stocks of the Euro Stoxx 50 and given the fact that I will use daily data. As suggested from early studies, the latter feature could be more valuable for the analysis [196], [197].

3.2 Central Banks' Policies and Herding behaviour

In recent years, it seems emerged the need to take into account the behavioural elements of market participant attitudes (in particular their psychological and cognitive biases [82]) when policy makers, market supervisors and regulators have to take policy decisions [181]. For instance, in a recent Consultation Paper, the European Securities and Markets Autohority (ESMA) illustrates how investors may be affected from heuristics in their decision making process that (in turn) could lead to sub-optimal outcomes[2]. Yet, Khan [181] describes how behavioural elements are relevant to financial supervision, regulation, and central banking and that their effective realization strictly depends on their comprehension (and integration) at policy makers level. In conclusion, it is worth noting the new strand of research related to the *Behavioural Monetary Policy Making* [105], [106], and, more in general, linked to the *Behavioural Macroeconomics*

[2]Consultation Paper E.S.M.A. 13 July 2017 | 35-43-748, p.9.

[89], [100], [153] both attempt to construct a new kind of models over the next few years, useful for policy decisions at the macro level.

However, when it comes to analysing herding behaviour and its destabilizing effects on financial markets, the literature that studies such a type of relationship is extremely scarce [186]. At the same time, it is in undeniable the relevance of such a pattern for central banks, especially given the "new roles" attributed to them in the post-2008 financial crisis [222]. Among others, the most relevant (as highlighted in section 2.2.2) is the one related to financial stability and (as we noted in section 3.1.1) *intentional* herding could the cause of systemic risk [16]. In this case, the role of central banks would be clearly exacerbated.

Hence, following these two important points, in section 3.2.1 we will see at the (few) papers that try to detect a relationship between central bank interventions and herding in financial markets, whilst in section 3.2.2 we will discuss the proposals advanced so far in order to address this kind of behaviour under a regulatory point of view.

3.2.1 Central banks and herding behaviour in financial markets

In order to provide a literature review of results regarding herding behaviour in response to central bank statements, it would be interesting to follow the same order illustrated in section 2.2.2. This would allow us to link a *specific* central bank activity (or macroeconomic release) to herding events observed in financial markets. However, it is worth noting that actually there seem to be studies that attempt to analyse these events only with respect to *monetary policy* decisions, thus leaving aside central banks' (possible) influence on releases concerning financial stability and economic outlook. With respect to the former, the FED's role would be of particular relevance, given its greater importance for the real economy with respect to other central banks [299].

Regarding *monetary policy*, as one would expect, first studies attempted to analyse its relative effects on the herding behaviour in the *bond* market. In a recent paper, Galariotis et al. [115] employed daily prices of 10 Government Benchmark Bond Indices for different countries inside the European Union between 2007 and 2011. In order to test for herd behaviour, they augment eq. (3.8) to take into account days during which fundamental macroeconomic information (from the ECB, Bank of England, and Federal Reserrve) was released. They found that during the EU crisis period, macroeconomic information induced bond market investor herding. Their evidence, in particular, supports the concept of *spurious* herding, since it was caused by changes in fundamentals (see section 3.1.1). In a similar work, Chirinko and Curran [67] examine the relationship between speeches, testimonies, and FOMC meetings (or STFs) and volatility in the 30-year U.S. Treasury bond futures market. By using intraday data, they found that STFs is an important predictor for bond market volatility, thus reinforcing the idea of spurious herding advanced from Galariotis et al. [115]. It is also worth noting that the same result is found for the Bank of Japan: in particular, Kamada and Miura [176] point out that central bank communication is, *in general*, an important driver behind such a type of behaviour. Moreover, they also highlight the importance of setting it as part of the strategy behind a good conduct of monetary policy.

In more recent years, the focus moved to analyse monetary policy impact on the *equity* market. Solakoğlu et al. [303] evaluate the effect of the central banks'

meetings (in particular from the central bank of Turkey, the ECB, the FED, the Bank of England and the Bank of Japan) on herding behaviour for Borsa Istanbul National Market 30 (BIST30) and Borsa Istanbul Second National Market (SNM). By using the techniques illustrated in eq. (3.3) and (3.4 - 3.7) they found evidence of herding on both the two types of market. However, some central bank (as one would expect) exert more influence than others: in their case, herding behaviour is detected for a post-meeting day for ECB, BOE and BOJ. Gond and Dai [124] investigate the effects of monetary policy announcements on herding in the Chinese stock market, using the same approach of Solakoğlu et al. [303]. They show that an announcement of raising the benchmark deposit rate leads to herding behaviour, but a similar effect is not observed for the opposite case. Lastly, in a more recent work, Krokida et al. [186] examine directly the relationship between conventional and unconventional ECB and USA monetary policy and herd behaviour in US and EU equity markets. The test for herding is pursued by means of the beta herding procedure proposed in eq. (3.9). They show that both central banks exercise an important influence on their respective equity market, but, however, the spill over effect (i.e. the impact that seemingly unrelated events in one nation, but that can have on the economies of other nations) is only attributable to FED policies. Notably, this result confirms the one observed earlier in Fig. 2.9, highlighting the unique impact that the FED exerts at global (*stock*) market level.

3.2.2 First attempts of anti-herding regulation

In section 3.1.1, we noted that in some cases, (in particular spurious) herding could be the cause behind systemic risk [318]. The same idea was highlighted, under a regulatory point of view, from Ayres and Mitts [16]. In particular, they look at herding as a negative externality that policy makers should try to tackle in order to prevent the economy from such a type of risk. Yet, other papers highlight the concern that rigid corporate regulation (in particular in the banking sector) and uncritical adoption of innovations may force firms into *similar* decisions that are *micro*-functional, but dysfunctional at the *macro* level [119], [133]. In particular, if one can identify a proxy of market stress that can significantly influence investors' behaviour, regulators can focus on those stress proxies in order to monitor market volatility and develop safety nets and circuit breakers [57]. Clearly, this is an important consideration as enlarged understanding of the herding-risk proxy relationship may help prevent the destabilizing effects of investor herding.

Moreover, as the 2008 credit financial crisis [141], [293], and, in a similar fashion, the subsequent European sovereign debt crisis in 2011 [39], [274] shed light, systemic risk could be the cause, in turn, of asset price bubbles [53]. It would then be of extremely importance to introduce some mechanisms to boost separating equilibria in order to avoid the flattening among the market consensus [16], [119], [318].

Hence, the idea proposed from the different papers introduced earlier it is to provide a general framework for policy makers in order to apply an *anti-herding regulation* [16]. It is beyond the scope of this Thesis to provide a full description of the different types of regulation (with related advantages and disadvantages) that attempts to tackle herding in the financial markets and across different sectors [83]. Here, we will only describe the most important insights that could be of interest for central bank purposes.

According to Ayres and Mitts [16], anti-herding regulation can produce two kinds of benefits, in particular for the banking sector.

The first one is that it can reduce the kinds of systemic risk that occur when there is excessive behavioural uniformity. For instance, excessive (and regulatory induced) clustering among the balance sheets of a relatively small number of banks can expose the whole financial system to the risk that a decline in one firm's assets will prompt a cascading wave of insolvency in the sector. In general, regulation is a natural response to the problem of negative externalities. At the same time, in some contexts (like the one described above), it is the pooling of behaviour that is itself the problem.

The second benefit is that anti-herding can produce socially beneficially information. In particular, by inducing *different* equilibria among several regulated entities, this could avoid the inefficiency of informational cascades (see section 3.1.1) and help to steer both private and public actors toward better evidence-based outcomes.

However, in spite of this, there not seems to exist a strand of literature (or a supranational regulatory report) that attempts to study anti-herding regulation looking at central banks as main author behind that. Instead, more attention is paid at how central banks' monetary policy should tackle asset price bubbles *ex-post* [128], [279], [233], or contagion [62] leaving aside what they could actually do in order to prevent herding behaviour in financial markets on an *ex-ante* basis.

For sure, the "regulatory recipe" proposed by Ayres and Mitts [16] and, together with them, Gerding [119] and Haiss [133], is ambitious and would require a systematic effort between different authorities in order to actually apply such a robust project. For instance, at the European level, this would require a joint work between the Single Supervisory Mechanism [110] (of which the ECB has the responsibility) and the different Authorities within the European System of Financial Supervision [253].

At the same moment, as we said in the introduction of this section, in the following years we can expect an increasing relevance of behavioural elements to financial supervision, regulation, and central banking [181]. In particular, the new "toolbox" that policy makers will have at hand will span from models that apply a behavioural approach to macro-finance from a complex systems perspective [153] to models that take into account how populism (and more in general, the crowd's sentiment) could affect central bank policies [107] and financial stability [347].

To conclude, and to link the relevance of central bank communication and herding behaviour in in financial markets, it is worth noting what Prast [265] states: "*An implication (of the fact that it is difficult to base policy prescriptions on the theory underlying herding behaviour of investors) may be that policy makers who care for socially optimal investor behaviour and, more importantly perhaps, financial stability, should engage in* ***extensive***, ***timely*** *and* ***careful information dissemination***".

Chapter 4

Case Study: Text Mining of E.C.B. Press Conferences

> *"What matters for transparency is therefore clarity as well as openness. For a new and supranational institution like the ECB, it is particularly important that it* ***sends clear and coherent messages to the markets and the wider public****."*
>
> - Otmar Issing, *Executive Board Member of the ECB - 1999*

> *"[...] by slightly increasing the price of leverage at an early stage of a developing boom, the central bank could break* ***herding behaviour*** *when the development of a bubble depends on investors observing other investors purchasing the bubble-prone asset."*
>
> - Lucas Papademos, *Vice President of the ECB - 2009*

This Thesis (and in particular this chapter) aims to quantifying communication of the European Central Bank (ECB), during the press conferences on the Governing Council[1] meeting days.

As we pointed out in section 2.2, a growing body of economic literature applies tools from computational linguistics to analyse central bank *communication*. It is worth stressing again the reason for this: communication became a key tool for central banks to maintain transparency [261], manage market expectations [73] and achieve policy goals in a zero-lower bound environment [45], [298].

Regarding the ECB, it is important to notice that nowadays in order to foster its transparency and "*to send clear and coherent messages to the markets*", it uses various channels to communicate its monetary policy stances [182]: (i) press conferences, (ii) monetary policy accounts, (iii) monthly bulletins, (iv) speeches, and (v) interviews. However, the *press conferences* (that take place on the same day as the Governing

[1]The Governing Council is one (together with the General Council and the Executive Council) of the three main ECB's bodies. Governing Council is the most important body because it is empowered to: (i) formulate the monetary policy within the Economic and Monetary Union (EMU); (ii) decide about the interest rates; (iii) manage the liquidity; (iv) deliberate on the reserves and the estimate of the ECB capital; (v) decides on the other tasks entrusted to the European System of Central Banks, the system formed from the ECB as well as the National Central Banks.

Council decision announcement) are the most important communication device. In particular, they are one of the most valuable sources of information for financial markets since they provide explanations about monetary policy decisions, financial stability opinions and economic outlook. The utility behind the usage of such a type of (text) data in text mining applications is threefold. First, given the fact press conferences are perceived as (the main) communication tool since 06/1998 (i.e., the first time in which an ECB press conference was held), an evaluation of their usefulness is now in order [101]. Second, as Schmeling and Wagner [288] suggest, since press conferences take place during trading hours, this implies that investors can react to new information instantaneously, and the staggered timing of rate announcement and press conferences allows one to disentangle market reactions to news about policy rates and communication. Third, this kind of information is sent to "*the wider public*". This means that the ECB opinions about specific topics regarding the state of the economy can affect the investor decision making process. In other words, it can happen that ECB *sentiment* could drive *herding* behaviour in the financial (in particular European) market, a pattern already showed in Fig. 2.1b.

In this chapter I will then combine the material described in chapters 1, 2 and 3 to analyse 123 press conferences of the European Central Bank, during the period 2008-2019. Specifically, the focus of the Thesis is to study how ECB's press conferences *sentiment* (or tone) can impact on the European stock market. The analysis will thus follow two different (but interconnected) stages. First, ECB press conferences will be extracted using a Web Scraping algorithm and then quantified by means of a Sentiment analysis procedure, comparing different field specific dictionaries [76], [207], [261], intrinsic to *this* type of communication. The resultant sentiment time series will be then compared to the Euro Stoxx 50 market realizations. The last goal is to understand whether Euro Stoxx 50 values can be affected from press conferences tone *on the same day* of ECB President announcements. Second, I will assess whether herding behaviour can be detected across different European financial sectors using an augmented form of eq. (3.7). Specifically, that equation will be constructed with the aim to take into account for both negative and positive tones related to ECB press conference, as well as resulting from the values of the several sentiment time series. The last goal is to analyse if ECB sentiment can affect investor herding behaviour across several European stock sectors. Results will show that ECB press conferences tone can improve our understanding of *both* these two kinds of phenomena.

The structure of this chapter is as follows: section 4.1 introduces the ECB press conferences as well as the web mining process in order to extract them and the financial data. Section 4.2 explains the text preprocessing steps applied to ECB communication and provides an Exploratory Data Analysis of the: (i) ECB's President speech and Q&A section during time; (ii) the application of different field specific dictionaries in order to construct several sentiment time series derived from the ECB's press conferences; and (iii) the $CSAD_t$ statistics (see section 3.1.2) of different European equity sectors. Section 4.3 discusses the empirical results and section 4.4 concludes this chapter.

4.1 Data

This section introduces the ECB press conferences and describes the preprocessing phases to convert ECB text statements to numerical data (in the same vein of section 1.5). It also presents the financial data used to measure the market reaction to the

ECB's sentiment during the press conferences announcement days.

4.1.1 ECB Press Conferences

In order to describe the retrieval of ECB press conferences data we will first describe what a Press Conference is, its structure and why it could provide important information to the stock market. Then, the Web Scraping algorithm (and its related computational features) to extract them from the ECB's website will be discussed.

1. ECB Press Conferences and meaning of each section

The ECB's monetary policy decisions are published at 13 : 45 CET on the day of the Governing Council monetary policy meeting. The press conference starts at 14 : 30 on the same day. It begins with an introductory statement of the ECB President who explains the monetary policy decisions. Before 2015, the Government Council meeting moved from twice in a month to one each 6 weeks.

The press conference consists of six major sections: (i) summary of the ECB's monetary policy decision (since July 2013 it includes also a forward guidance); (ii) economic analysis; (iii) monetary analysis; (iv) "cross-check" paragraph; (v) fiscal policy and structural reforms; (vi) questions-and-answers (Q&A) section. Sections (i-v) belong to the ECB's President *introductory statements*.

The monetary and the economic analysis are the two pillars by which the Governing Council evaluates the risk to price stability[2]. The economic analysis part looks at short to medium-term outlook, whilst the monetary analysis assesses medium to long-term trend[3]. The *cross-check* paragraph was introduced in 2003 and its role is to compare signals from the two pillars.

My work will analyse all ECB press conferences between January 2008 and October 2019, covering 46 speeches from Jean-Claude Trichet (whose eight-year term expired at the end of October 2011), and 77 speeches from Mario Draghi (whose mandated expired at the end of October 2019). The textual data has been scraped from the ECB website[4]. While historical statements are available from the 06/1998, the choice of starting point (i.e. January 2008) was made to strike a balance between "quantity" and "quality" (adequate representation through different periods of the economic cycle such as recession (2009), sovereign debt crisis (2011) and subsequent, slow but steady,

[2]A clarification here is important. According to the Maastricht Treaty (ex. art 105), the monetary policy objective of the ECB is the maintenance of the *price stability*. The ECB clarified that price stability must be meant as an increment in the Harmonised Indices of Consumer Prices (HICP), for the *whole* Eurozone, below or near to the 2%. *Without prejudice to the objective of price stability*, the ECB shall also sustain other general economic policies in the European Union, with a view to contributing to the achievement of the objectives of the EU (ex. art 2) [139]. These (often called secondary) economic policies comprehend: (i) a harmonious and balanced development of economic activities; (ii) a sustainable growth (iii) a low level of unemployment.

[3]The difference between the two horizons is due to the ECB "monetary policy strategy oriented to the overall stability" [222]. It is formed from two pilasters: (i) The first one is (actually) the *inflation targeting* and belongs to the economic analysis. It implies a general assessment of the outlook for Euro area prices. In particular, this valuation aims at identifying the risks for price stability to the *brief and medium term*: it is then grounded on the analysis of macroeconomic trends, as well as the shocks that affect the system and risks to the prices stability; (ii) The second one is the *monetary targeting*: it tends to analyse the trends in the in the inflation in a *medium-long term*. It is referred, in particular, to the tendencies of the monetary aggregates, and a benchmark value for the growth rates in these aggregates are announced periodically.

[4]`https://www.ecb.europa.eu/press/pressconf`

growth in equity market values). Hence, there is a total of 123 statements published during this period.

2. Preparing Documents by means of a Web Scraping algorithm

In my work, for each press conference, I developed an algorithm to break the ECB's President introductory statement from the Q&A section. The reason to treat these two sections separately is that the former has a *standardized* structure (as confirmed from Amaya and Filbien [8]), whilst the latter does not. A possible "omission" of this feature could substantially impact the final model output (as suggested in several text mining studies [93], [165]).

Moreover, text similarity is also important for the Web Scraping process (see section 1.4.7). In this thesis, such a process has been applied using the `rvest` package (see section 1.6.3). To give a graphical representation of the introductory statement similarity (and to understand how my algorithm actually works) we can look at Fig. 4.1, which represents the press conference of the 24/10/2019 (and available here[5]). In particular, as we can see from Fig.4.1a, each press conference (web) document begins with a link to redirect the user to the Q&A section, while in Fig. 4.1b it is showed how generally the ECB's president concludes his introductory statements. These two features allow us to create two different *regular expressions*, also known as *regex*. As we explained in section 1.4.7, regex are strings with a special syntax that allow one to match patterns in other strings. Hence, via the function `read_html()` we allow `R` to connect to the ECB's website, whereas via the function `str_locate()` we identify where these two regex actually are in the document.

An introductory statement section will be thus be defined as *one vector of multiple strings*: specifically, it will be the "interval" that goes from the string following the regex "*Jump to the next transcript*" up to the regex "*at your disposal*". Instead, the Q&A section will be defined as the interval that goes from the string following the regex "*at your disposal*" up to the end of the document.

This procedure was then applied to each one of the 123 press conferences. However, it could happen that in some specific cases one of the two regex described earlier was missing in a document. To check this, I then created a logical vector indicating whether (and in which specific document) that condition was not respected. There was a specific case in which this happened. On the press conference held on the 8/12/2011[6] the introductory statement ended with the sentence "*Product market reforms should focus on fully opening up markets to increased competition.*". That specific document was then treated a part.

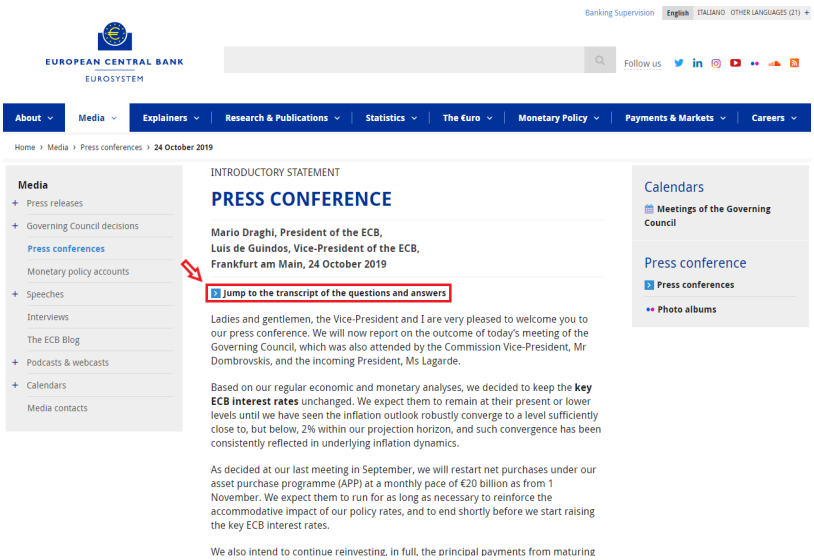
Beside text data scraped from ECB press conferences, I also extracted other useful information such as: (i) the statement date; (ii) the relative year of that press release; (iii) the statement length. These three other kinds of information (defined to be meta-data, see section 1.1.1), will be particularly useful in the exploratory data analysis and in the construction of the sentiment time series.

Thus, at the end of the `for` cycle, a `data.frame` object will be created, containing the following columns:

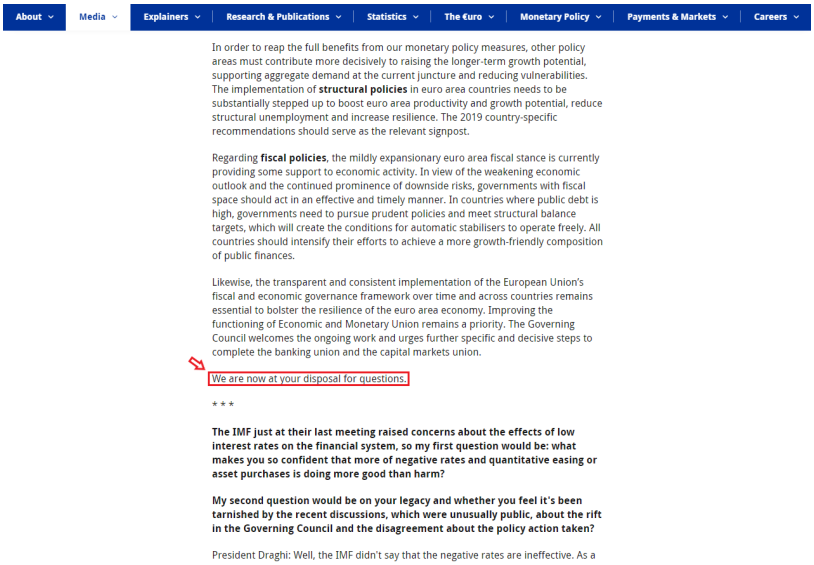
- The year of the statement,
- The statement date,

[5] `https://www.ecb.europa.eu/press/pressconf/2019/html/ecb.is191024~78a5550bc1.en.html#qa`

[6] `https://www.ecb.europa.eu/press/pressconf/2011/html/is111208.en.html`



(a) How a general introductory statement *begins*



(b) How a general introductory statement *ends*

Figure 4.1: A representation of a general ECB press conference (web) page.

- The URL of that particular press release,
- The statement content (i.e. the vector of multiple strings defined early),
- The statement length.

We can think at this `data.frame` object as a *Corpus* (see section 1.5.2): each additional (Web Scraped) document implies new information about ECB communication, as arising from its press conferences.

4.1.2 European Financial and Sector data

This Thesis uses the *Euro Stoxx 50* index (hereafter STOXX50) to study the effect of changes in ECB tone on Eurozone equity returns on press conference days. The STOXX50 is then used as a proxy of the European stock market returns. This (in general) seems to be a reasonable assumption since it covers the 50 largest firms in the

Eurozone [27], [52]. Moreover, this index has also been also investigated in the context of ECB communication and monetary policy statement by Picault and Renault [261] and from Schmeling and Wagner [288], in order to measure (in both cases) the impact from ECB press conferences tone on the European stock market. The sample period for the STOXX50 mirrors the one of the press conferences (i.e. from January 2008 to October 2019) with 3,033 daily observations, of which 123 are press conferences days (with tone changes) and 2,910 are non-press conferences days. Data have been retrieved from Thomson Reuters.

Additionally, in order to analyse the possibility of herding behaviour around the different component stocks of the STOXX50, sectors data are used. Specifically, I retrieved data for the 50 component stocks of the STOXX50 from Thomson Reuters for the same length of period of the STOXX50 (and thus of the press conference sample). The division of stocks according to their relative sector is showed in table 4.1.

Sector	Firms	
Energy&Raw materials	Air Liquide	Eni
	BASF	Iberdrola
	CRH	Linde PLC
	Enel	Total
	Engie	Vinci
Financial	Allianz	ING Groep N.V.
	AXA	Intesa Sanpaolo
	B.B.V.A.	Munchener Ruck AG
	BNP Paribas	Banco Santander
	Deutsche Börse	Société générale
Healthcare	Bayer AG	
	EssilorLuxottica	
	Fresenius SE	
	Philips	
	Sanofi	
Industrial Goods&Services	Airbus Group	Schneider Electric
	B.M.W.	Siemens AG
	Daimler	Volkswagen
	Deutsche Post DHL	
	Safran	
Personal&Household Goods	Adidas	Inditex
	Ahold Delhaize	Kering
	Amadeus IT Group	L'Oréal
	Anheuser-Busch InBev	LVMH
	Danone	Unilever N.V.
Technology&Comm.	ASML Holding	Telefónica
	Deutsche Telekom	Vivendi
	Nokia	
	Orange S.A.	
	SAP	

Table 4.1: Component stocks of the Euro Stoxx 50 and relative sectors.

In order to calculate $CSAD_t$ values (see eq. 3.4) for each sector, daily log-returns are determined using the relation $R_{i,t} = ln(\frac{P_{i,t}}{P_{i,t-1}})$ where $P_{i,t}$ represents the daily closing prices of day t for the stock i. I then calculated the returns of the market portfolio based on equally weighted portfolios of all firms in each sector classification according to the taxonomy provided in table 4.1. Moreover, I also calculated $CSAD_t$ values for an equally weighted portfolio formed from the benchmark indices related to the same sectors presented in table 4.1. For instance, for the financial sector, the Euro STOXX Banks EUR Price Index and the EURO STOXX Insurance EUR Price Index are both considered to be one of the $R_{i,t}$ returns used to calculate the final equally weighted portfolio. This portfolio will be called "EU sector portfolio" (or ESP). A possible detection of herd behaviour in such a portfolio would imply that herding does not cancel out also in the case of sectoral diversification.

4.2 ECB text preprocessing and Exploratory Data Analysis

Through the Web Scraping process illustrated in the previous section we created a Corpus of documents related to the ECB press conferences. However, in order to prepare our data for Sentiment Analysis we have to transform them from unstructured to structured, as explained in section 1.5.

Hence, in subsection 4.2.1 we will explain how the text mining preprocessing steps are applied to the ECB press conferences, describing also how the different packages introduced in section 1.6.3 were used. In subsection 4.2.2, instead, we will look at the several dictionaries used to "filter" such a type of documents and the relative construction of the different sentiment time series.

Moreover, in subsection 4.2.3 we will look at descriptive statistics for the equally weighted portfolios as well as the $CSAD_t$ values related to different European sectors.

4.2.1 ECB's President Introductory Statements and Q&A sessions

Before to apply the text preprocessing techniques we will first explore the information provided by the metadata. Then, we will structure our data exploring *how* ECB communication changed over time. We will try also to understand the different reasons behind such a topical change.

1. Exploratory Data Analysis using metadata from ECB Press conferences

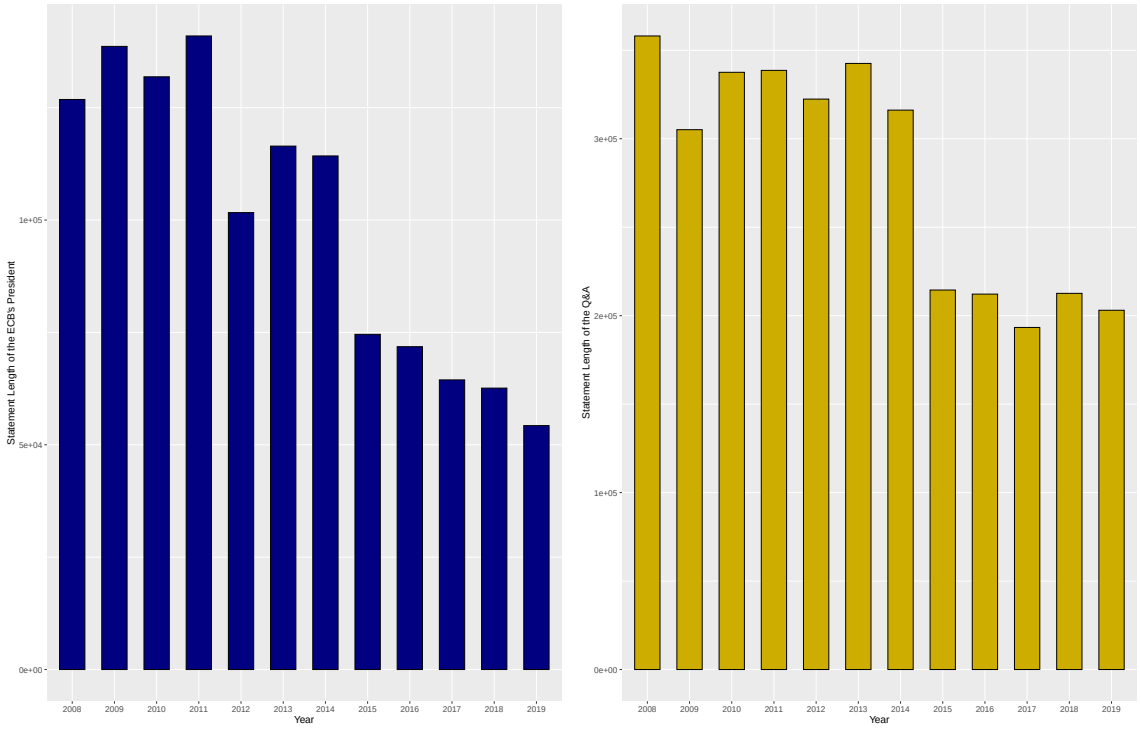
Figure 4.2 shows how much the total statement length of press conferences per year changed during time, dividing for the introductory statement and Q&A section. As we can see, both pictures exhibit a downward trend. We could link this pattern to the different events that conditioned the ECB mandate during the period 2008-2011, where the two major crisis (one global and the other more at European level) happened. At a first glance, we could say that in periods of higher financial distress for the whole market, the ECB communication strategy tends to be more "prolix". Moreover, also the Q&A section tends to be more "verbose", maybe in order to attempt to give an answer to market concerns in these kinds of periods.

However, to stress this hypothesis, it would be more interesting to see how the total statement length per year changed according to the ECB's President. This is showed in Fig. 4.3, where the introductory statement (Fig. 4.3a) and Q&A section (Fig. 4.3b) are compared according to the ECB's President duration mandate. This picture reinforces our hypothesis made for Fig. 4.2, since the ECB press conferences became progressively verbose (and together with them, the Q&A section) under the Trichet mandate, and reached they peak in 2011, when Mario Draghi took over as the new ECB's President.

Notably, a similar pattern is detected in the Data Article from Warin and Senger [330] which gathers all press conferences in text from made during the period 1998-2016. This also confirms the good task achievement of my Web Scraping algorithm.

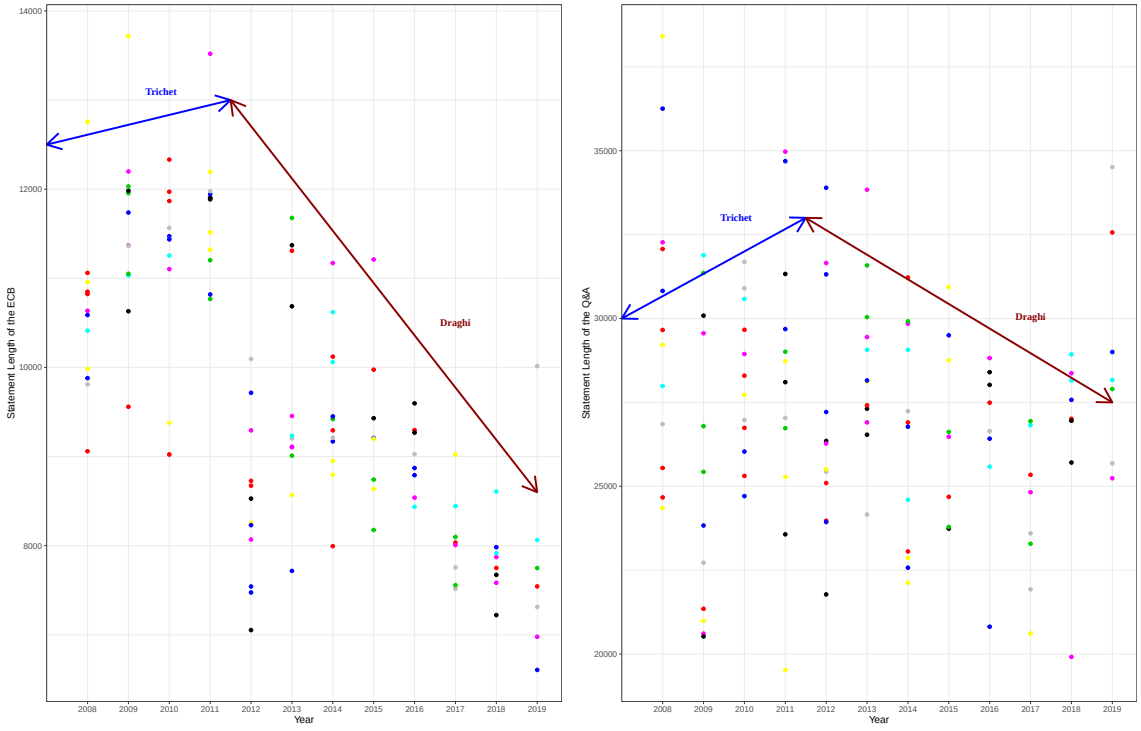
2. ECB communication, term weights and EU financial history

In order to get new insights from ECB press conferences (like for instance how communication changed over time) we have to structure our data.



(a) Statement Length of the Introductory statement part **(b)** Statement Length of the Q&A section

Figure 4.2: Total statement length of press conferences per year.



(a) Statement Length of the Introductory statement part **(b)** Statement Length of the Q&A section

Figure 4.3: Total statement length of press conferences per year according to the ECB's President

First, I create a list of stopwords related to the neutral sentences that are repeated in every speech and don't provide useful information for text mining analysis. Examples of such sentences are: "*Ladies and gentlemen, the Vice President and I are very pleased to come you to our press conference*"; "*Let me now explain our assessment in*

greater detail, starting with the economic analysis"; "*We are now at your disposal for questions*". A similar procedure is applied to the Q&A section. The complete list of expressions that were removed is provided in Appendix B.

Secondly, I convert all words to lower case, remove numbers and punctuation, and identify other stopwords within the document by means of the `tm` package. I also use the stopwords provided in the package `tidytext`. To perform the first three tasks the `tm` package was used, whereas the latter operation was applied by means of the `anti_join()` function.

Third, I identify collocations and create uni-grams, in line with the bag of word assumption introduced in section 1.5.2. The `unnest_tokens()` allow us to select a specific document and to tokenize words in uni-grams. In spite of the fact that such hypothesis is criticized in different papers (related also to ECB central bank communication [182], [261]) the main reason to apply such a strategy is related to the fact the two dictionaries that I will use to construct my "central bank" lexicon (and that will be introduced in section 4.2.2), rely on uni-grams rather than on n-grams.

Notably, tasks in step two and three can be executed in `R` at the same time by means of the pipe operator (or "%>%") as we can see from the code below:

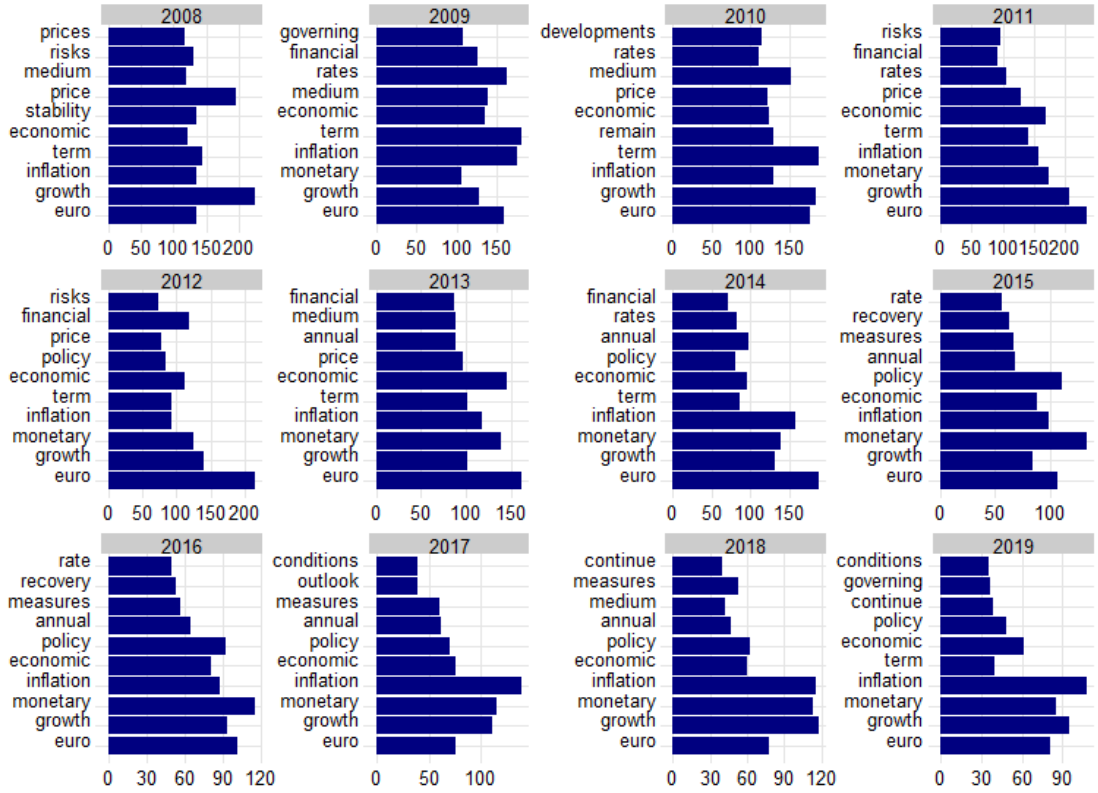
```
report.words.ecb <- reports_ecb %>%
  mutate(date = statement.dates, year = year, text =
     statement.content.ecb) %>%
  unnest_tokens(word, text) %>%
  mutate(word = stripWhitespace(gsub("[^A-Za-z ]","", word))
     ) %>%
  filter(word != "") %>%
  filter(word != " ") %>%
  anti_join(new.stop_words.ecb) %>%
  count(date, year, word, sort = TRUE) %>%
  mutate(frequency = n) %>%
  select(date, year, word, frequency)
```

Lastly, I reordered tokens each year and ranked them according to their (absolute) frequency. I also calculate the *tf-idf* values, introduced in eq. (1.5), to get more insights about the differences in ECB communication during time. For both the press conferences and Q&A section, *tf-idf* values have been multiplied times 100 to allow for a better interpretability.

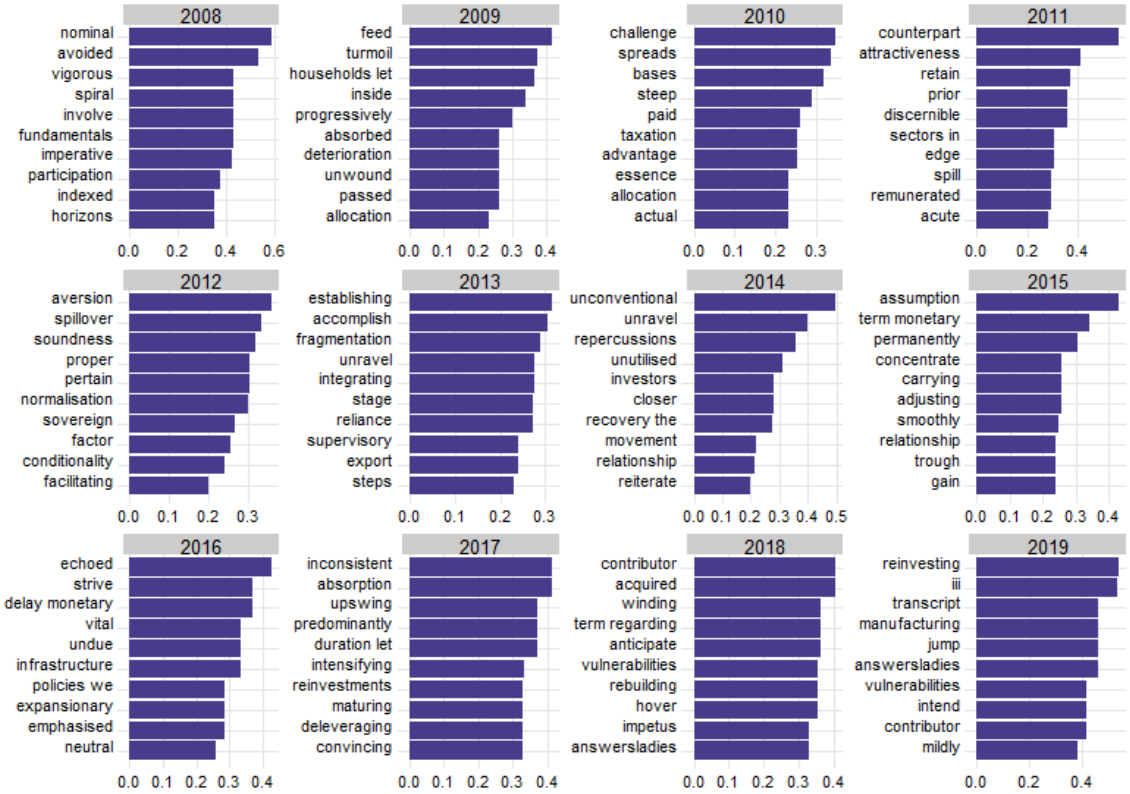
The final result can be seen in Fig. 4.4 and 4.5 for the introductory statement and Q&A section, respectively. Both figures give a good impression of the breadth of issues that the ECB communication was concerned with, and how it changed over time.

It is then interesting to link these major changes in ECB communication during years, as showed in Fig. (4.4 - 4.5), to the most important events that this central bank had to face during the period 2008-2019, and the related regulatory changes that affected its mandate within the European Union. The description of such a "timeline" of events will be important also to explain the reasons behind the decision to pick up specific dictionaries for Sentiment Analysis. Yet, it will offer also the possibility to understand certain patterns about the values of the sentiment time series. A chronology of events regarding the Eurozone debt crisis (and subsequent events in

Figure 4.4: Total statement length of press conferences per year (Introductory statement part).

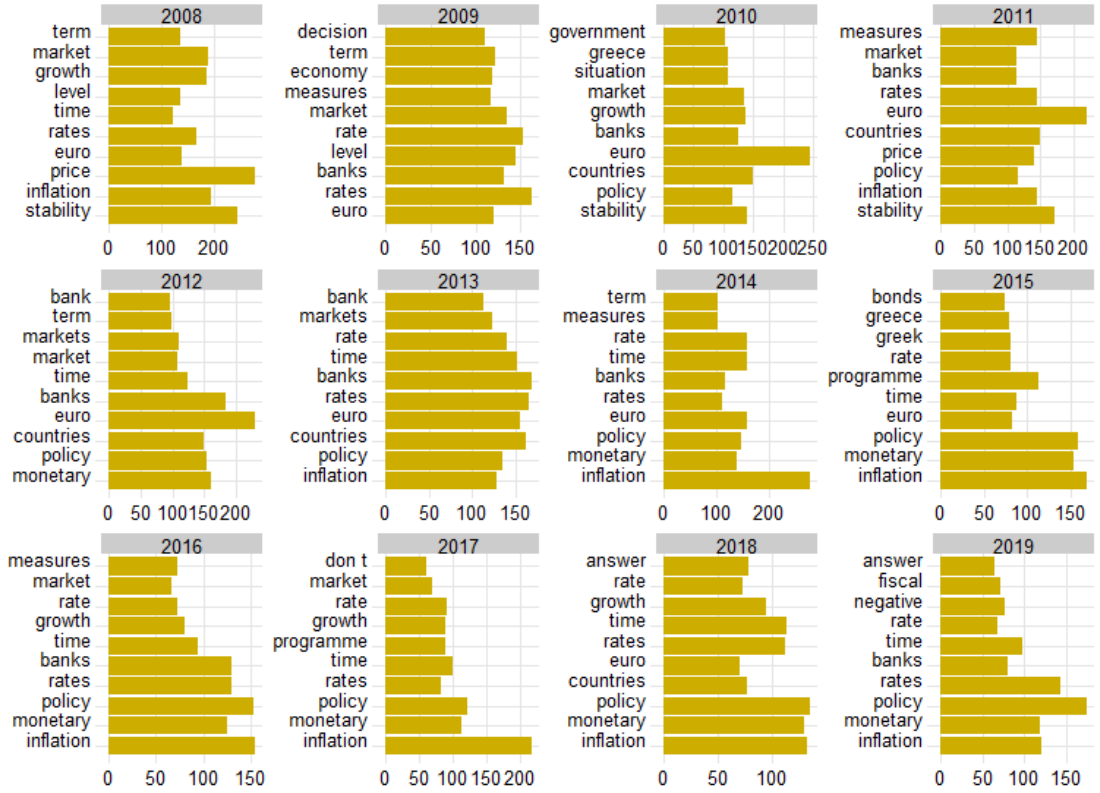


(a) Most frequent uni-grams for the introductory statement part.

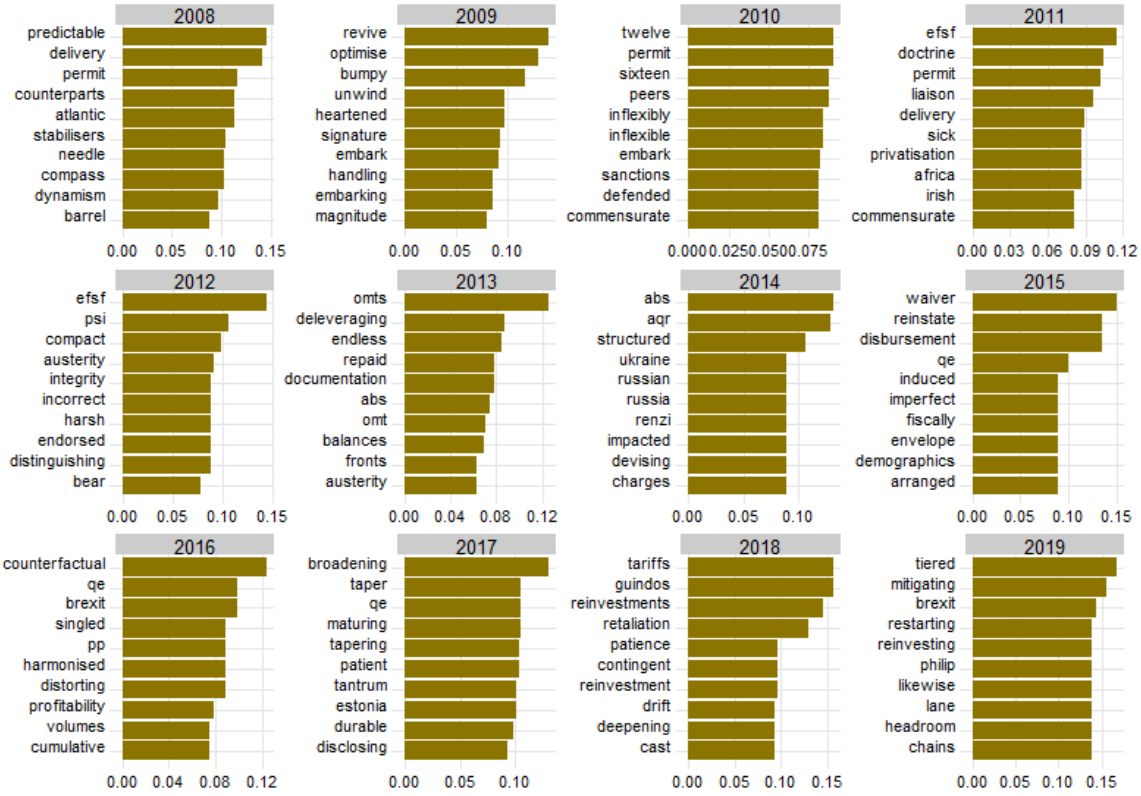


(b) Tokens with the highest *tf-idf* values for the introductory statement part.

Figure 4.5: Total statement length of press conferences per year (Q&A section).



(a) Most frequent uni-grams for the Q&A section.



(b) Tokens with the highest *tf-idf* values for the Q&A section.

the following years) can be found on the BBC[7] and IG[8] websites. Instead, the major challenges in the ECB monetary policy mandate can be found in the (ECB) working paper of Hartmann and Smets [139] and in the work of Rodriguez and Carrasco [276].

My sample period starts from the 10/01/2008, the same year of Lehman Brothers bankrupt. At the outset of the economic crisis in 2008, the EU mirrored the rescue packages seen in the US: more than €1.6 trillion was used to save the banks and €200 billion to help boost European growth following the global financial crisis. However, this huge financial aid didn't seem enough for some 'peripheral' EU country, such as Greece whose bonds were downgraded (from Moody's, Standard&Poor's and Fitch) over worries about its ability to reduce its high debt levels (that reached 113% of its GDP). The ECB communication in that (and following) years was in line with a response to market concerns and the tools that it used to face such a crisis. In particular, the words "*term*", "*rates*", "*medium*", "*allocation*" and "*absorbed*" all refers to the huge effort of the ECB to help economy (such as the Long-Term Refinancing Operations (LTROs), the Covered Bond Purchase Programme 1 and 2 (CBPP1 and CBPP2), the so-called Outright Monetary Transactions (OMT) programme (that we can find also in Fig. 4.4b), among others). Also the Q&A section followed a similar pattern, where the terms "*stability*", "*rates*", "*banks*", "*measures*" and "*greece*[9]" were the most questioned topics. Beside the normal press conferences, the overall communication strategy was clearly aimed at calming financial markets. The most famous "example" is the now iconic sentence by Draghi during a generic conference held in London: "*Within our mandate, the ECB is ready to do* ***whatever it takes*** *to preserve the Euro. And believe me, it will be enough*".

However, the sustained financial and economic crisis led new important regulatory changes staring from 2010. First, the common opinion that the financial supervisory framework did not enough to prevent all the different scenarios observed in previous years drove Finance Ministers to the creation of the European Financial Stability Facility (ESFS). The same term appears for the first time in 2011 in Fig. 4.5b, the same year in which the ESFS entered into force. Second, the creation of the Banking Union was another important step to ensure financial stability to the whole EU system. The Banking Union relies on three pillars: (i) the Single Supervisory Mechanism (SSM); (ii) the Single Resolution Mechanism (SRM); and the (iii) European Deposit Insurance Scheme. The SSM, as defined from the SSRM ex art. 2 it is: "*the system of financial supervision composed by the* ***ECB*** *and national competent authorities of participating Member States*". The ECB was then officially chosen to pursue supervision of institution in their jurisdictions in order to foster *financial stability* tasks, beside its role of monetary policy [92]. It is in fact since the 2013 that terms like "*supervisory*", "*reliance*" and "*investors*" entered into the ECB (introductory statement) jargon, as we can see from Fig. 4.4b. Hence, also if in subsection 4.1.1 we explained that in a generic press conference there is no a specific part that addresses only the financial stability, it would be misleading not to take into account such a "topical shift" in ECB communication [182]. This is why (among other dictionaries) a financial stability lexicon, as provided in the work from Correa et al. [76] was used in order to analyse ECB communication tone.

Analysing both Fig. 4.4 and 4.5 for the following years, we can detect other interesting information in ECB press conferences, in particular with respect to the

[7] `https://www.bbc.com/news/business-13856580`

[8] `https://www.dailyfx.com/research/eurozone-debt-crisis#info`

[9] In Fig. 4.5b, the term "Greece" appears without the capital letter because of the text mining preprocessing phases described early.

Q&A section. For instance, the ECB's President were required to give his opinion about topics such as: (i) Governments' austerity measures[10]; (ii) Geopolitics issues related to Ukraine and Russia[11]; (iii) Brexit[12]; (iv) Asset backed securities[13] and Quantitative Easing measures[14].

4.2.2 ECB's Sentiment and the European equity market

The joint study of the timeline describing the main financial events in EU history, combined with the changes in ECB communication, allowed us to analyse all the different topical features behind ECB press releases [182]. The aim of this subsection is thus to find the means by which such a communication can be *numerically* quantified.

In particular, having cleaned and structured data at hand, we can actually apply one of the different text mining models introduced in section 1.5.4. In this thesis, *Sentiment Analysis* will be applied to 123 press conferences during the period 2008-2019. Sentiment Analysis allows one to extract the polarity of the expressed opinion in a range spanning from positive to negative tone, using a dictionary that allows to *filter* (that is, to give a *numerical* value) to textual data. Hence, in this subsection I will explain all the different dictionaries that were used to quantify ECB communication and the analytical process to perform such a task, both for the introductory statement and for the Q&A section.

To link the importance of measuring ECB press conferences tone during (European) stock trading days, as Schmeling and Wagner [288] suggest, it is worth noting the fact that press conferences take place during trading hours, implying that investors can react to new information instantaneously, and the staggered timing of rate announcement and press conferences allows one to disentangle market reactions to news about policy rates and other (topic-related) communication. In other words, one could expect that *equity markets respond to changes in tone in ECB press conferences*, as already showed in Fig. 2.1b. In this work, the *standardized* values of several sentiment time series will be then compared to the *standardized* values of the STOXX50 realizations during the period spanning from the 2008 to 2019. Importantly, in this subsection we will analyse this relationship in a graphical way, while in subsection 4.3.1 we will look at the statistical implications of such an association.

In `R`, there are two main packages in order to apply Sentiment Analysis: the first one is `SentimentAnalysis`, whereas the second is `sentometrics`.

The former fosters the Sentiment Analysis process when the dictionary is "*binary*" (i.e. is formed only from positive and negative words). The way in which sentiment is applied at document level (here represented from a generic ECB press conference) is straightforward. In particular, given a document q as an input, the `analyzeSentiment()` function computes the sentiment of a specific document as follows:

$$Sentiment(q) = \frac{PositiveWords - NegativeWords}{PositiveWords + NegativeWords} \tag{4.1}$$

[10] `https://www.ecb.europa.eu/events/pdf/conferences/141215/papers/MUELLER_Austerity.pdf?0b98714766958c064ae1b18457da0d19`

[11] `https://www.ecb.europa.eu/pub/financial-stability/fsr/focus/2014/pdf/ecb~9151d8b64f.fsrbox201405_03.pdf`

[12] `https://www.ecb.europa.eu/pub/fie/article/html/ecb.fieart202003_01~690a86d168.en.html`

[13] `https://www.ecb.europa.eu/mopo/implement/omt/html/abspp-faq.en.html`

[14] `https://www.ecb.europa.eu/explainers/show-me/html/app_infographic.en.html`

In our case, the final results will be a (single) score for each one of the 123 press conferences during the period that spans from 2008 to 2019.

The latter, instead, is useful when the specific lexicon is created allowing for a *weighting scheme* (i.e., not all words in a document level receive the same weight in order to measure the overall tone). In their (reference manual) paper, Ardia et. al [12] define the sentiment measure as:

$$s^l_{n,t} = \sum_{i=1}^{I} w_i v_i s^l_{i,n,t} \tag{4.2}$$

The sentiment of a document is thus given by the sum of the confidence scores $s^l_{i,n,t}$ for each word in the document which is included in the lexicon l. Instead, w_i is the within-document aggregation weight and v_i denotes the valence-shifters words (i.e., words that increase, decrease or change the polarity of a sentence). As authors suggest, in the case of an uni-gram approach, this measure can be interpreted as a weighted sum of sentiment-scores for all words within a document which are listed in a lexicon.

It is worth stressing that the standardized form of ECB statements during years (empirically showed by the uni-grams used in in Fig. 4.4a) implies that the application of a non-field specific lexicon may fail to capture all specificities of central bank communication, as already explained in section 2.2.1. Albeit standardized, however, ECB communication does not refer only to monetary policy. Hence, in my work, five different dictionaries are used in order to measure ECB press conferences tones: (i) the Loughran-McDonald (henceforth LM) dictionary [207]; (ii) a Financial Stability (hereafter FS) dictionary, introduced in the work of Correa et al. [76]; (iii) my dictionary (MD), that simply arises from the combination of the LM and FS dictionary (and thus follows their binary structure); (iv) a monetary policy and (v) an economic outlook (EC) (weighted) dictionaries, both taken from the work of Picault and Renault [261]. Given this framework, the dictionaries (i-iii) are applied to ECB press conferences by means of the `SentimentAnalysis` package, whilst dictionaries (iv-v) by means of the `sentometrics` package.

The reasons to pick out these dictionaries are in line with developments in ECB communication during years and with previous work [182], [261], [288].

The LM dictionary (introduced in section 2.1.3), despite its wide usage in a Corporate Finance domain, has been also used in central bank communication context, and in particular with respect to ECB press conferences tone [288]. In particular, Schmeling and Wagner [288] show that this lexicon captures important features of ECB communication, such as macroeconomic fundamentals (a result also confirmed from Picault and Renault [261]).

The other three dictionaries were all introduced in section 2.2.1. The FS dictionary has been constructed using financial stability reports coming from a panel of 35 countries [76]. Comparing the LM and FS dictionary, Correa et al. find that over 30% of the positive or negative words in their dictionary are not classified in LM's dictionary. Thus, by applying only LM dictionary to central bank communication one could lose relevant information. To the best of my knowledge, this is the first work that apply a financial stability lexicon to ECB press conferences. As explained in the previous subsection, this is justified from the new role of financial stability supervisor attributed to the ECB with the creation of the SSM [92].

The (uni-gram version) of the MP and EC dictionaries have been downloaded

from the Picault and Renault website[15]. The choice to select these two dictionaries is related to the fact that they are constructed by means of an algorithm applied *directly* to the ECB introductory statements. In order to compare the performance of MD dictionary with the MP and EC dictionaries, their uni-gram version will be considered.

In conclusion, it is important to notice that since no lexicons have been found for the Q&A sections, I used the same dictionaries introduced earlier for the introductory statement part.

In order to allow comparability between the STOXX50 and sentiment time series values, both the two time series have been standardized. Formally, the daily values for the STOXX50 and for the other sentiment time series (both related to the introductory statement and to the Q&A section) will be defined as:

$$z_{i,t} = \frac{x_{i,t} - \mu_i}{\sigma_i} \tag{4.3}$$

where μ_i and σ_i are, respectively, the in-sample mean and in-sample standard deviation of the specific time series of interest, for the period that spans from 01/2008 to 10/2019. With respect to the time interval t, the dates between STOXX50 and ECB press conferences were uniformed by means of the function `inner_join()`, in the `dplyr` package.

The final results are shown in Fig. 4.6 and 4.7[16] for the introductory statement and Q&A sections, respectively. Moreover, table 4.2 exhibits in-sample correlations between the STOXX50 and the relative score of a specific dictionary, both for the introductory statement and Q&A sections.

Dictionaries	$\rho_{I.S.,STOXX50}$	$\rho_{Q\&A,STOXX50}$
L.M.	0.319	0.128
F.S.	0.483	0.070
M.P.	0.006	0.133
E.C.	0.541	0.118
M.D.	0.545	0.038

Table 4.2: In-sample (Pearson) correlation coefficients between STOXX50 values and the sentiment analysis applied to the Introductory Statement (I.S.) and Question and Answers (Q&A) section.

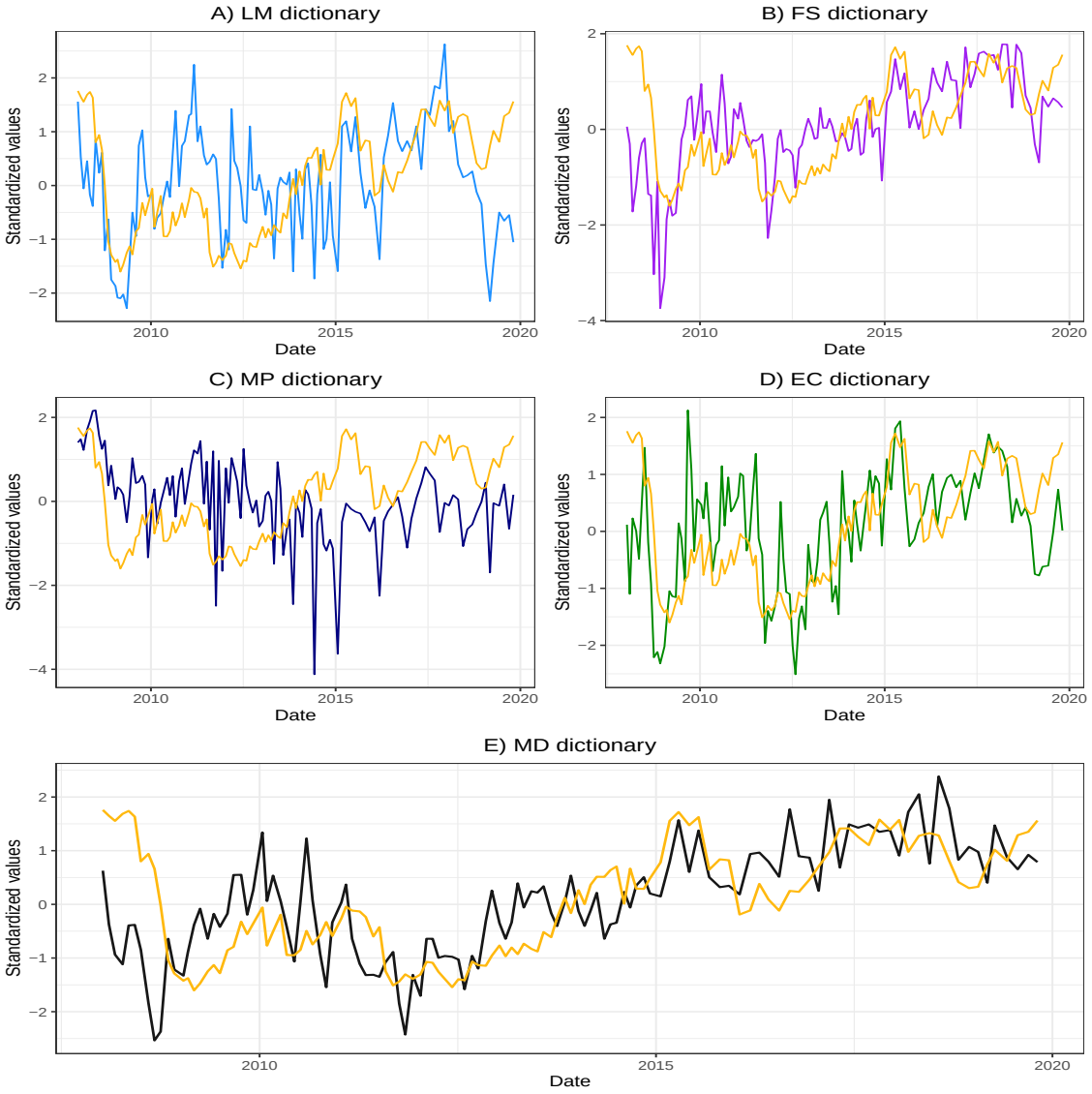
As it clearly emerges from a rapid examination of Fig. 4.6 and 4.7, the different dictionaries applied to the introductory statement part allow us to track changes in the STOXX50 during time in a better way than those applied to the Q&A section. We will then pay more attention to the former ones.

Specifically, all dictionaries seem to adapt well in the first part of the sample, but after the 2010 results seem to diverge, not providing a significant relationship with the STOXX50 series anymore. This happens in particular for the LM and MP dictionaries. On the other hand, the EC dictionary seems to track well the STOXX50 movements, both in downwards and upwards periods. Moreover, its positive and negative spikes seem also to *predict* similar (directional) movements in the stock market index. The same can be said for the FS index, also if in the middle part of the sample period the EC seems to perform better. However, the MD dictionary improves the FS behaviour

[15] `http://www.cbcomindex.com/`

[16] It should be noted that, in both Fig. 4.6 and 4.7 *continuous* values are obtained during time by means of a linear interpolation [111].

Figure 4.6: Standardized values between sentiment analysis applied to the I.S. part and STOXX50.

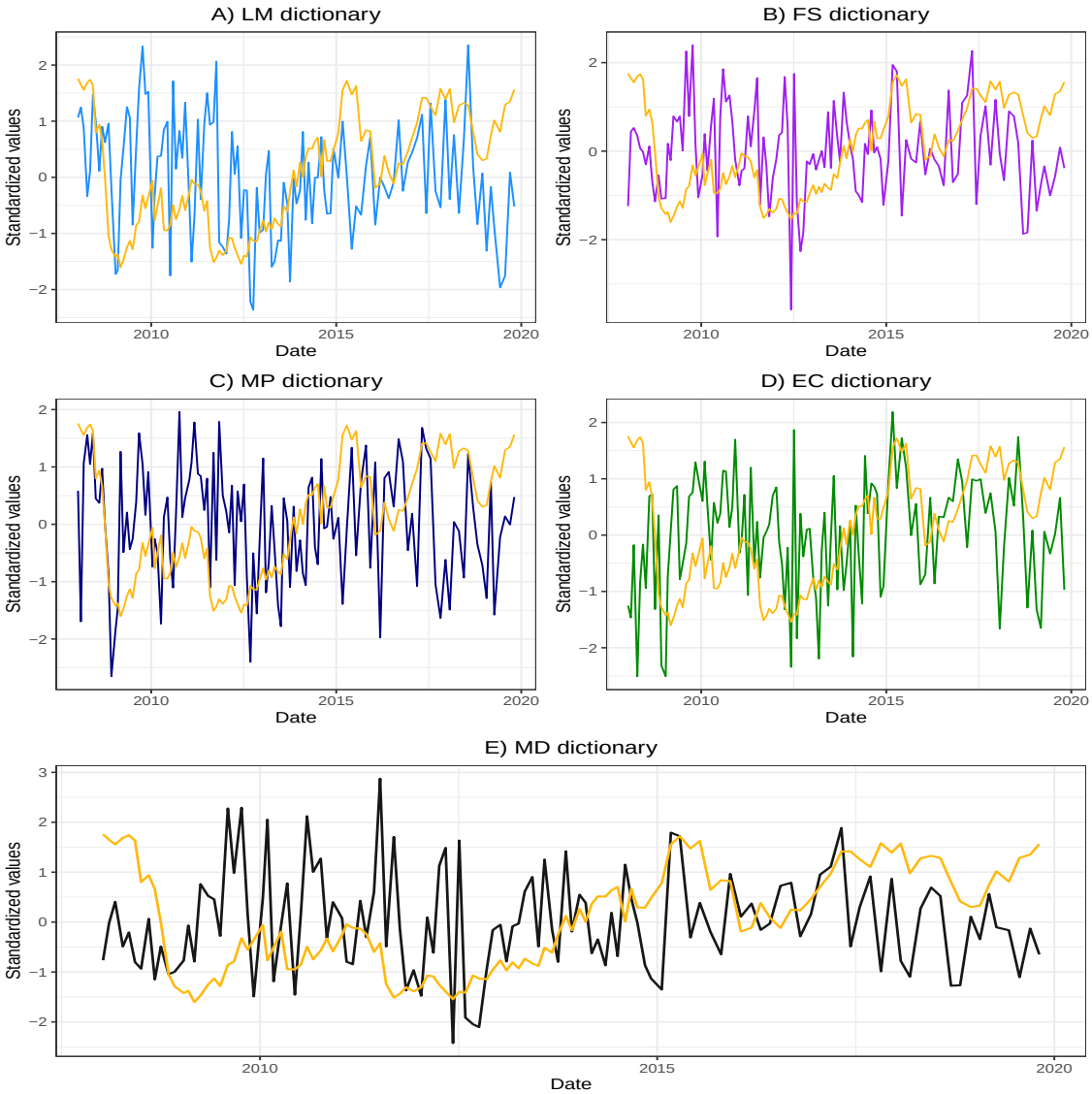


*Notes: The figure exhibits the standardized values of the STOXX50 series (yellow) together with five different sentiment time series related to the ECB's President introductory statement part: (i) the LM series (light-blue) in Panel A; (ii) the FS series (magenta) in Panel B; (iii) the MP series (blue) in Panel C; (iv) the EC series (green) in Panel D; and (v) the MD series (**black**) in Panel E. All series have been standardized by means of eq. (4.3).*

by integrating it with the LM dictionary. We can interpret this result in a qualitative way. First, a reason for this improvement in tracking STOXX50 realizations could be attributable to the fact that the financial stability role of the ECB started after the 2012 (i.e. after the creation of the Banking Union), and thus its communication moved towards this topical contents only after that year. The market was thus embedding this kind of information "inside" the STOXX50 values only after that period. The second reason for such an improvement could be related to the intrinsic structure of the LM dictionary, that makes it particularly suitable in a stock market context. The LM lexicon was in fact constructed on 10-Ks: we can thus speculate about the fact that there could be a lot of words in ECB communication that could affect investors in this segment, and (in turn), move European asset prices.

To conclude, it is interesting to assess in a qualitative way whether the EC and MD sentiment time series track well the different events introduced in section 4.2.1 and those that could be of particular interest to the European stock market. As we

Figure 4.7: Standardized values between sentiment analysis applied to the Q&A section and STOXX50.



*Notes: The figure exhibits the standardized values of the STOXX50 series (yellow) together with five different sentiment time series related to the Q&A section: (i) the LM series (light-blue) in Panel A; (ii) the FS series (magenta) in Panel B; (iii) the MP series (blue) in Panel C; (iv) the EC series (green) in Panel D; and (v) the MD series (**black**) in Panel E. All series have been standardized by means of eq. (4.3).*

can see from Fig. C.1 (where the EC and MD series are shown in Fig. C.1a and C.1b, respectively) it seems that they did. In particular, even if they do not always show a (quantitative) significant relationship with the STOXX50, their positive and negative spikes allow us to detect important events related to the ECB financial history. Regarding the EC dictionary, it seems that it plunged during financial concerns about the Greek debt, while it recovered according to future (positive) ECB projections about the whole EU economy. Notably, it seems to drop again given "*[...] the persistence of uncertainties related to [...] vulnerabilities in emerging markets (that) appears to be leaving marks on economic* ***sentiment***"[17]. The same can be said for the MD dictionary, which dropped due to serious concerns about the stability of the whole EU financial system, and rose in times of important programmes launched by the ECB,

[17] `https://www.ecb.europa.eu/pub/economic-bulletin/html/eb201902~a070c3a338.en.html`

or thanks to the good financial performances of this central bank.

4.2.3 Financial and Herding statistics for the European stock market

Table 4.3 summarizes the descriptive statistics for average daily returns of market portfolios, while table 4.4 reports similar information about $CSAD_t$ values for the different sectors considered in this work. Moreover, a graphical representation of $CSAD_t$ values of different European sectors is shown in Fig. D.1. Panel G of Fig. D.1 refers to the CSAD values of the ESP portfolio.

With respect to table 4.3, we reveal that all sectors are characterized by average daily returns, standard deviations and medians consistently low, a result in line with Ouarda et al. [249]. The high Kurtosis value for the Energy sector is maybe related to the crisis of the European Electricity System[18].

Table 4.3: Descriptive statistics of the average daily returns for different European sectors

Sector	Mean	Median	Max.	Min.	Standard Dev.	Skewness	Kurtosis
Energy&Raw materials	0.00%	0.02%	11.76%	-11.67%	0.74%	0.47	102.08
Financial	-0.01%	0.01%	6.96%	-7.34%	0.89%	0.08	8.32
Healthcare	0.01%	0.02%	3.67%	-3.14%	0.54%	-0.08	3.17
Industrial Goods&Services	0.01%	0.03%	4.97%	-4.13%	0.70%	-0.22	4.71
Personal&Household Goods	0.01%	0.03%	3.72%	-4.77%	0.53%	-0.27	6.78
Technology&Comm.	0.00%	0.01%	3.91%	-3.39%	0.56%	0.06	4.43
E.S.P.	0.00%	0.01%	4.05%	-3.49%	0.55%	-0.14	5.50

Notes: The table presents descriptive statistics for daily log returns of equally weighted portfolio related to the ESP and to six European financial sector, including: (i) Energy&Raw materials; (ii) Financial; (iii) Healthcare; (iv) Industrial Goods&Services; (v) Personal&Household Goods; and (vi) Technology&Comm. All data are available at Thomson Reuters and range from 01/2008 to 10/2019. Missing information for holidays was carefully inspected or interpolated.

Examining table 4.4, we notice abnormal Skewness and Kurtosis values for the Energy&Raw material, Personal&Household Goods, and the whole EU sectors. In this case the outcomes of such statistics could be attributable to the different outliers showed in Fig. D.1.

Table 4.4: Descriptive statistics of $CSAD_t$ statistics for different European sectors

Sector	Mean	Median	Max.	Min.	Standard Dev.	Skewness	Kurtosis
Energy&Raw materials	0.42%	0.33%	20.78%	0.07%	0.89%	20.656	449.866
Financial	0.38%	0.31%	2.74%	0.07%	0.26%	2.589	10.691
Healthcare	0.34%	0.28%	2.81%	0.02%	0.23%	2.513	11.203
Industrial Goods&Services	0.38%	0.32%	3.12%	0.04%	0.24%	3.061	18.145
Personal&Household Goods	0.36%	0.31%	10.22%	0.07%	0.27%	17.504	607.363
Technology&Comm.	0.37%	0.31%	3.09%	0.06%	0.23%	2.938	17.881
E.S.P.	0.25%	0.22%	2.53%	0.00%	0.14%	4.911	53.478

Notes: This table lists descriptive statistics of daily, equally weighted cross-sectional absolute deviations ($CSAD_t$) for the ESP and six European financial sector, including: (i) Energy&Raw materials; (ii) Financial; (iii) Healthcare; (iv) Industrial Goods&Services; (v) Personal&Household Goods; and (vi) Technology&Comm. All data are available at Thomson Reuters and range from 01/2008 to 10/2019. Missing information for holidays was carefully inspected or interpolated. $CSAD_t$ values were calculated according to eq. (3.4).

4.3 Results

This section analyses the European stock market reaction to ECB press conferences in two different ways.

[18] https://www.strategie.gouv.fr/sites/strategie.gouv.fr/files/atoms/files/CGSP_Report_European_Electricity_System_030220141.pdf

First, I assess whether ECB communication can explain the evolution of the STOXX50 on the same statement days. Specifically, each sentiment time series (both for the introductory statement and Q&A section) will be regressed on the standardized values of the STOXX50 during the period 01/2008 to 10/2019. In the same vein of Picault and Renault [261], and Schmeling and Wagner [288], the aim is to analyse *which* components of ECB communication, if any, can impact the European stock market.

Second, I consider if negative and positive values in ECB communication tone can be one of the drivers behind herding behaviour across different sectors in the European stock market.

4.3.1 E.C.B. Sentiment Analysis, Asset Pricing and Cointegration

In a different way from Picault and Renault [261] and Schmeling and Wagner [288], in my Thesis I will use the "raw" values of the STOXX50, and not its (log) returns. A stylized fact in Finance is that many financial time series (such as the one of the STOXX50) do not exhibit stationarity: in other words, their statistical properties, such as mean, volatility and covariances are *not* constant over time. However, financial time series are often made stationary in order to ease the statistical process to get *parsimonious* models. At the same time, by applying a stationary technique to the original (or raw) time series, one can lose *relevant* information embedded inside the series. This is ever truer in a *multivariate* time series context. Specifically, in the case of two or more non-stationary processes, the Economic and Financial theory states that a long run equilibrium relationship could exists around which these two series move together. It could then happen that a *linear* combination of integrated variables is stationary, a case known as *cointegration* [122]. In Economics, cointegration is important to estimate the long run equilibrium relationships among macroeconomic variables. In Finance, cointegration can be used to find trading strategies based on mean-reversion, a phenomenon also known as *statistical arbitrage* [280].

Hence, to detect the possibility of cointegration between two or more economic variables one has to follow a two-step process. First, one must be sure that all the time series of interested are nonstationary (or, more formally an integrated of order one, or I(1), process). Second, one has to analyse whether the residuals of the regression of the dependent variable $Y(t)$ on $X(t)$ are stationary, in order to prevent the possibility of *spurious* regression. In this subsection this process will be applied to understand whether ECB communication tone is cointegrated with the standardized values of the STOXX50.

In Statistics, in order to tell whether a time series is stationary or not stationary, one can use hypothesis testing by means of *unit root tests*. In general, unit root tests are used to decide if an autoregressive model has an absolute root equal to 1 (i.e. the condition for the stationarity of the process is violated). Three popular unit root tests are the (i) augmented Dickey-Fuller test (or ADF test) [65]; (ii) the Phillips-Perron (hereafter PP) test [260]; and (iii) the KPSS test [149]. In the first two tests the null hypothesis is that there is a unit root, whereas the KPSS test follows an opposite logic. To apply these three tests in `R`, one can use the `adf.test()`, `pp.test()` and `kpss.test()`, respectively. These three tests are applied to both the sentiment time series arising from the introductory statement and Q&A section. Also if sometimes these test tend to give contrasting results (not showed) there is the overall confirmation

that all the series in Fig. 4.6 and 4.7 are not-stationary.

Hence, the standardized values of the STOXX50 are regressed on the standardized values of each of the sentiment time series introduced in section 4.2.2. For all regressions, the MD dictionary results are compared with the outcomes arising from the LM, FS, MP and EC sentiment time series. Specifically, to explain the link between different ECB press conferences tone and the European stock market on the *same* day of the press conference, I considered the following regression model:

$$z_t^{STOXX50} = \beta_i^l z_t^l + \varepsilon_t \tag{4.4}$$

where $z_t^{STOXX50}$ represents the variation of the STOXX50 on the same day of a generic press conference and β_i^l is the i^{th} regression coefficient of the l lexicon. On the press conference day (and given the previous results from Picault and Renault [261] and Schmeling and Wagner [288]) I expect β_i^l to be positive for the LM, FS, EC and MD dictionary, as good (bad) news in a corporate, financial stability, and economic contexts should improve (decrease) companies' rationally discounted future cash flows. On the other hand, I expect β_i^l to be negative if the MP indicator incorporates information about future monetary policy stances, in the same vein of Picault and Renault [261].

Tables 4.5 and 4.6 present results for the introductory statement and Q&A section, respectively.

Table 4.5: Contemporaneous relationship regression results between sentiment time series applied to the I.S. and STOXX50.

	(1)	(2)	(3)	(4)	(5)
(Intercept)	0	0	0	0	0
	(0.09)	(0.08)	(0.09)	(0.08)	(0.08)
z_t^{LM}	0.32 ***				
	(0.09)				
z_t^{FS}		0.48 ***			
		(0.08)			
z_t^{MP}			0.01		
			(0.09)		
z_t^{EC}				0.54 ***	
				(0.08)	
z_t^{MD}					0.55 ***
					(0.08)
N	123	123	123	123	123
R-squared	0.1	0.23	0	0.29	0.3

*Notes: This table presents the results from the linear regression introduced in eq. (4.4) related to the introductory statement part. The dependent variable for all models (i.e. [1-5]) is the raw daily value of the STOXX50 on ECB statement days, or $z_t^{STOXX50}$. Sample from 01/2008 to 10/2019. Standard errors are reported in parenthesis. Significance at 1%, 5%, 10% are denoted respectively by ***,**,*.*

As we can see from regression results in table 4.5, it seems that (as found from Picault and Renault [261] and Schmeling and Wagner [288]) ECB introductory statement tone about corporate, financial stability and economic topics is *positively* associated with equity levels. This indicates that more (less) positive tone about these arguments imply, on average, higher (lower) values of the standardized series of the STOXX50 on the same day of the ECB press conference release. However, this cannot be said with respect to the MP time series. This result in contrast with the finding of Picault and Renault [261], indicating that the choice between an uni-gram or n-gram approach

Table 4.6: Contemporaneous relationship regression results between sentiment time series applied to the Q&A section and STOXX50.

	(1)	(2)	(3)	(4)	(5)
(Intercept)	0	0	0	0	0
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
z_t^{LM}	0.13				
	(0.09)				
z_t^{FS}		0.07			
		(0.09)			
z_t^{MP}			0.13		
			(0.09)		
z_t^{EC}				0.12	
				(0.09)	
z_t^{MD}					0.04
					(0.09)
N	123	123	123	123	123
R-squared	0.02	0	0.02	0.01	0

*Notes: This table presents the results from the linear regression introduced in eq. (4.4) related to the Q&A section. The dependent variable for all models (i.e. [1-5]) is the raw daily value of the STOXX50 on ECB statement days, or $z_t^{STOXX50}$. Sample from 01/2008 to 10/2019. Standard errors are reported in parenthesis. Significance at 1%, 5%, 10% are denoted respectively by ***,**,*.*

can significantly affect output results [182], [261]. Despite the fact that all other regressions exhibit a statistically significant value with the standardised STOXX50 series, the EC and FS series seem to provide best in-sample results. In particular, the EC and MD sentiment reached the highest R^2 in-sample outcomes. This finding could be attributable to the high impact that these kinds of topics can exert on the European stock market, in particular if they are pronounced by an Institutional body such as the ECB.

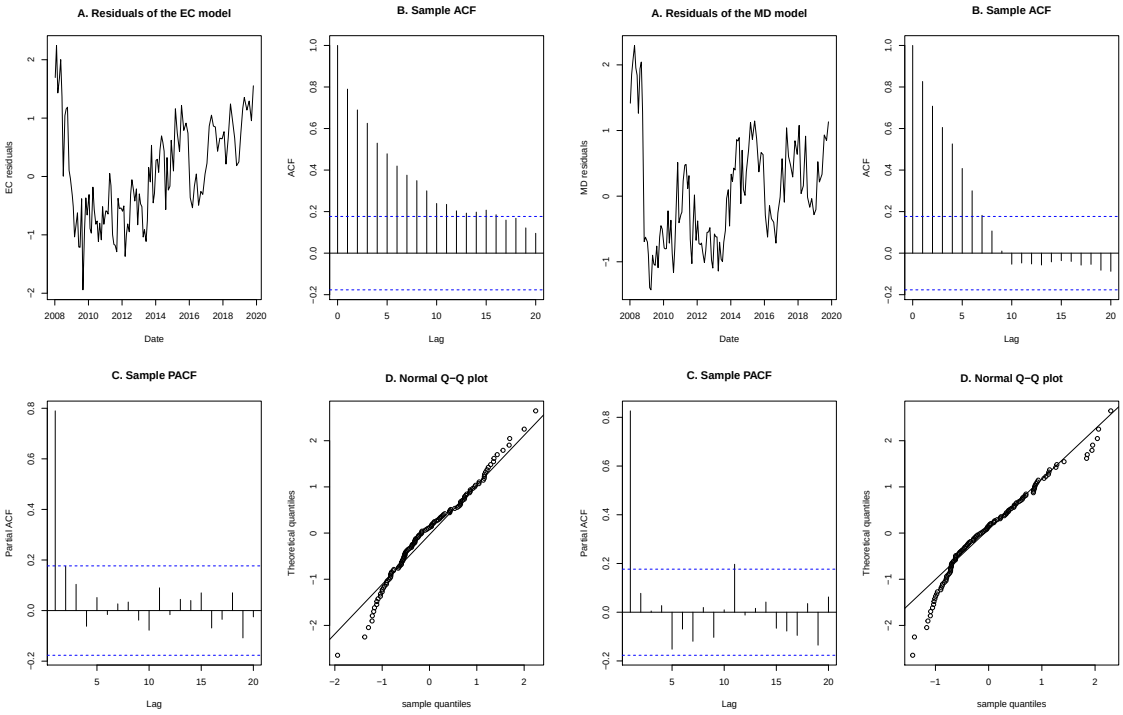
On the other hand, results related to the Q&A section do not show a similar pattern. In particular, *none* of the regression coefficients have a statistical relationship with the $z_t^{STOXX50}$ series. Albeit Fig. 2.1b showed that the European stock market reaction happened *during* the Q&A of that particular day, the less standardized topical style of this type of communication is, in general, more difficult to capture for an automated dictionary process. Overall, this issue lowers the quality (and the meaning) of the Sentiment Analysis process.

Coming back to the analysis of table 4.5, it should be noted that these results may be confuted, as we stated early, in the case of *spurious* regression [126]. The problem here arises with respect to the residuals of a generic regression model where two or more time series are involved: in the extreme case where the residuals were an I(1) process, the least square estimator (the same used to estimate eq. 4.4) would be inconsistent, meaning that it will not converge to the true parameter as the sample size increases [280]. However, this does not happen if two or more series are cointegrated, since in such a case the *linear* combination of I(1) variables is a stationary process [122]. The key point here is to understand that if $Y(t)$ (here represented from $z_t^{STOXX50}$) is regressed on $X(t)$ (here represented from z_t^l) and the two series are cointegrated, then the residuals will be I(0) so that the lest-squares estimator will be *consistent*. More importantly, cointegrated time series will be suitable for regression analysis. The errors of such a regression model could show signs of *autocorrelation*, but there is a solution to this problem: one can replace the assumption of independent noise by the *weaker* assumption that the noise process is stationary, but possibly correlated. This

approach is referred to as an autoregressive moving average X (or ARMAX) models [22], in which the X indicates the inclusion of exogenous regression variables.

Now, we will first look at the residuals diagnostic of the EC and MD model in order to have a general understanding of the goodness of their regression outputs. Then, in order to test for cointegration between the different sentiment time series and the STOXX50, a Phillips-Ouliaris [259] cointegration test will be used.

Residuals diagnostic for the EC and MD models are shown in Fig. 4.8a and 4.8b, respectively. Panel A and Panel B of both figures plot the residuals of the two sentiment time series. The residuals of the EC model do not seem a stationary process (at least in the first part of the sample) and its autocorrelation function (hereafter ACF) in Panel B decays to 0 slowly. This could be a sign of either nonstationary or possibly of stationary with long-memory dependence [21]. On the other hand, the ACF of the MD model decays to zero quite quickly, indicating that the MD series could be stationary. Moreover, looking at its partial autocorrelation function (or PACF), a suitable model for its residuals could be an autoregressive of order 1 (or AR(1)) process. In both cases, the plot of normal Q-Q plot (Panel D) seem to exhibit a concave pattern, indicating that the residuals could be right skewness (as compared to the normal distribution).



(a) Residuals analysis of the EC model

(b) Residuals analysis of the MD model

Figure 4.8: Residual diagnostics for the EC (4.8a) and MD (4.8b) regression model.

However, since the two residuals plots are ambiguous about whether the residuals are stationary, a test of cointegration will be applied. The Phillips-Ouliaris cointegration test regresses one integrated series on other and applies the Philipps-Perron unit root test (introduced early) [260] to the residuals. The null hypothesis is that the residuals are unit root nonstationary, which implies that the series are *not* cointegrated. In `R`, the Phillips-Ouliaris test can be run by means of the `po.test()` function, which belongs to the `tseries` package. The p-value is computed by interpolation if it is within the range of a table in Phillips and Ouliaris [259]. Applying this test to the

five different time series (only related to the introductory statement) one finds for the LM, FS and MP lexicons the message "`p-value greater than printed p-value`" so that the null of the test is not rejected. On the other hand, the p-values for EC and MD series are 0.07281 and 0.08796, respectively, indicating that the residuals are stationary and so (according to this test), the two series are cointegrated with the standardized values of the STOXX50.

These results allow us to pursue our analysis applying the aforesaid ARMAX model combined with the least square estimator used to estimate regressions in Table 4.5. Formally, eq. (4.4) now becomes:

$$z_t^{STOXX50} = z_t^{MD} + \xi_t \tag{4.5}$$

where:

$$(1-\phi_1 B)\xi_t = (1+\theta_1 B)u_t$$

and $u_1, \ldots, u_n$ is a white noise process [280].

The goal is then to find a suitable (and stationary) model for the residuals of the EC and MD models. In `R`, to have an idea of which model could be fitted to them, one can use the function `auto.arima()` that employs both the Akaike's information criterion (or AIC) [251] and the Bayesian information criterion (or BIC) [242] to discern across different models. When applied to the EC and MD series, we find that both the AIC and BIC values give us the same outcomes for the two different residual series. Specifically, both the AIC and BIC select an ARIMA(0,1,1) for the residuals of the EC model, and an ARIMA(1,0,0) for the residuals of the MD model. These results are important for two reasons. First, they confirm the possible nonstationarity concerns about the residuals of the MD model (as showed before in the ACF and PACF in Fig. 4.8). Second, they also corroborate our hypothesis that a stationary model is suitable for the residuals of my model (in particular an AR(1) model) and thus makes sense to pursue eq. (4.5).

The final model output is showed in table 4.7. We note that the coefficient of z_t^{MD} is now nearly equal to zero and no longer significant. Instead, the ϕ_1 (i.e. the coefficient of the AR(1) model) is more than 30 times its standard error, indicating strong statistical significance.

Table 4.7: Regression estimates of the ARMAX model for the MD sentiment time series

Coefficient	Estimate	Std. Err.	t value
(Intercept)	0.3882	0.5593	0.6941
z_t^{MD}	0.056	0.047	1.1915
ϕ_1	0.9508	0.0283	33.597

Notes: This table presents the results from the linear regression introduced in eq. (4.5) for the MD series. Residuals follow an AR(1) process. The dependent variable is the raw daily value of the STOXX50 on ECB statement days, or $z_t^{STOXX50}$. Sample from 01/2008 to 10/2019. The last column shows the t-values for the significance test related to the: (i) intercept of the MD model, (ii) the MD series z_t^{MD} coefficient; (iii) residual coefficient ϕ_1.

A natural question is why the z_t^{MD} coefficient becomes smaller and insignificant if the errors are assumed to be an AR(1) process in eq. (4.5). As explained in Ruppert [280], one possible reason could be due to the small sample size the analyst has at

hand. In this work 123 values are observed during the sample period. However, with more data it might be possible to separate the effects on the STOXX50 daily values of my dictionary and noise autocorrelation.

However, it is also worth noting the possible implications that a presence of cointegration between contemporaneous values of ECB tone and realizations of the STOXX50 could have for both economic and financial purposes. Regarding the former, we could speculate about the fact that, in the long run, European stock market values, as proxied from the STOXX50, mean-revert to the (average) ECB tone about corporate and financial stability topics. Regarding the latter, as we said before, cointegration can be used in Finance to find trading strategies based on mean-reversion. One can then use the *raw* values of these two time series for a *statistical arbitrage* on the same day of ECB press conferences statements. At the same time, it is important to bear in mind that, unlike *pure* arbitrage, statistical arbitrage is an opportunity where a profit is only likely, not guaranteed [280].

4.3.2 Herding Behaviour across European Sectors

The second aim of this Thesis is to understand whether ECB press releases tone could affect herding behaviour in the European stock market across sectors.

In order to test this, I will augment eq. (3.7) by taking into account the negative and positive values for the EC and MD dictionaries showed in Fig. 4.6). I chose these two dictionaries given their higher performance in tracking developments in change in ECB communication, as showed in table 4.2. Formally, eq. (3.7) becomes:

$$CSAD_t = \beta_0 + \beta_1 R_{m,t} + \beta_2 |R_{m,t}| + \beta_3 R_{m,t}^2 + \sum_{k=1}^{p} \beta_k D_k^{Pos} R_{m,t}^2 + \sum_{j=1}^{n} \beta_j D_j^{Neg} R_{m,t}^2 + \varepsilon_t \quad (4.6)$$

where:

- D_k^{Pos} and D_j^{Neg} are two dummy variables taking the value of 1 on the days of k^{th} positive and j^{th} negative ECB tone.
- The terms $\sum_{k=1}^{p} \beta_k D_k^{Pos} R_{m,t}^2$ and $\sum_{k=1}^{n} \beta_j D_j^{Neg} R_{m,t}^2$ allow us to detect herd behaviour around ECB positive and negative tone.

Statistically significant and *negative* values for β_k and β_j indicate that the relationship between $CSAD$ and $R_{m,t}^2$ is non-linear under the influence of ECB communication tone, which also reflects the herding behaviour in a specific European sector around these press conference announcements. Regarding the EC sentiment time series in Fig. 4.6, 37 negative and 86 positive sentiment values were observed. Similarly, for the MD series, 29 negative tone values have been detected, while 87 were positive. The resemblance between these two patterns could be attributable to the high correlation between the two sentiment time series (of 0.82, specifically). To the best of my knowledge, this is the first work that apply such a similar procedure to analyse the relationship between central bank *communication* and investor herding.

In the spirit of Chiang and Zheng [66], I estimated eq. (4.6) using a Newey-West consistent estimator [243]. The regression outputs for the EC and MD dictionary are showed in tables 4.8 and 4.9, respectively. Notably, we see that this procedure allows us (albeit with different results) to detect herd behaviour across different EU sectors around ECB press conferences days. More importantly, results see to be *asymmetrical*

within sectors. Specifically, it seems that investors in the Healthcare sector herd around negative statements about the economic outlook and positive tone about the financial stability of the system. Similarly, the whole European market seems to be affected from negative ECB's sentiment about the economic outlook. The latter outcome would be problematic for an investor that wants to hedge her portfolio again herding (that in turn could imply volatility [56], [86]) concerns. Other results (with a lower value of statistical significance, at the 10% specifically) can be observed for the Energy sector and the Financial sector, with the former being affected from negative tone about financial stability and corporate sentiment, and the latter from positive economic outlook. However, in both cases, the estimated coefficient is positive, not allowing us to draw a conclusion about the possibility of investor herd behaviour in these specific sectors.

Table 4.8: Estimates of herding behavior across European financial sector by means of eq. (4.6) using the EC series as further control variable.

Sector	(Intercept)	R_m	$\|R_m\|$	R_m^2	$D^{Pos}R_m^2$	$D^{Neg}R_m^2$
Energy	22.73 ***	0	0.20 ***	4.82 ***	0.32	-4.89
	(3.65)	(0.00)	(0.01)	(0.27)	(2.09)	(2.98)
Financial	17.36 ***	0	0.15 ***	3.54 ***	0.58⋆	-0.15
	(0.77)	(0.00)	(0.01)	(0.25)	(0.33)	(0.36)
Healthcare	40.93 **	0	0.14 ***	5.44 ***	-1.51	-10.93 ***
	(12.65)	(0.00)	(0.02)	(1.28)	(5.42)	(3.16)
Industrial	26.49 ***	0.01 **	0.09 ***	4.31 ***	-1.84	-2.52
	(3.58)	(0.00)	(0.01)	(0.59)	(2.97)	(1.92)
Personal	15.93	0.01	0.14 ***	5.60 ***	4.72	-3.58
	(8.44)	(0.00)	(0.02)	(0.73)	(7.97)	(2.75)
Technology	14.15	0	0.14 ***	3.51 ***	2.27	5.73
	(8.32)	(0.00)	(0.02)	(0.86)	(4.53)	(6.97)
E.S.P.	15.19 ***	0	0.11 ***	4.08 ***	3.33	-4.55 *
	(2.84)	(0.00)	(0.01)	(0.47)	(2.16)	(1.79)

*Notes: This table reports the regression results of $CSAD_t$ based on eq. (4.6), using the EC positive and negative tone values as further control variables in that equation. The data range is from 01/2008 to 10/2019. Numbers in the parentheses are standard errors. Standard errors are corrected for autocorrelation and heteroskedasticity using the Newey-West method [243]. Significance at 1%, 5%, 10%, 15% are denoted respectively by ***,**,*,⋆.*

Regarding the other control variables used in eq. (4.6), it is worth noting that results are in line with the ones of Ouarda et al. [249], also if they use a bigger sample size and monthly values for each sector. Moreover, also here the proof of herding observed for the rest of sectors underlines that the potential of this pattern differs across sectors (in terms of the scale of the several coefficients).

It is then now interesting to understand whether this kind of herding behaviour is *rational* or irrational.

Looking at the definition given in section 3.1.1 about *spurious* (and thus rational) herding, we can say that this pattern could be attributable to the fact that investors in certain sectors tend to adopt a similar decision making process around the ECB press conferences announcements, as they seem to have similar investment decisions when exposed to the *this* kind of information.

At the same time, it is also worth noting that investors tend to herd more when the economic and financial stability (and corporate sector) outlook is negative, a results found also in the work of Ouarda et al. [249]. This observation offers a circumstantial evidence (to the contributions of the behavioural finance) that herding behaviour is more likely to occur when the uncertainty in the market is higher. In particular, this could be due to the *loss aversion* characteristic of investors, as explained in the paper of Tverksky and Kahneman [326]. In other words, during trading days in which the

Table 4.9: Estimates of herding behaviour across European financial sectors by means of eq. (4.6) using the MD series as further control variable.

Sector	(Intercept)	R_m	$\|R_m\|$	R_m^2	$D^{Pos}R_m^2$	$D^{Neg}R_m^2$
Energy	17.73 ***	0	0.19 ***	4.90 ***	-3.25	3.66*
	(3.74)	(0.00)	(0.01)	(0.28)	(1.91)	(3.23)
Financial	17.27 ***	0	0.15 ***	3.57 ***	1.37	-0.87
	(2.32)	(0.00)	(0.01)	(0.24)	(1.59)	(1.59)
Healthcare	39.00 ***	0	0.15 ***	5.38 ***	-14.00 ***	-4.29
	(6.68)	(0.00)	(0.02)	(1.39)	(3.92)	(5.42)
Industrial	25.45 ***	0.01 **	0.09 ***	4.24 ***	-2.4	-0.86
	(3.75)	(0.00)	(0.01)	(0.63)	(2.24)	(2.98)
Personal	17.47 *	0.01	0.14 ***	5.71 ***	3.55	-4.06
	(8.88)	(0.00)	(0.02)	(0.82)	(8.45)	(2.64)
Technology	14.34	0	0.14 ***	3.56 ***	-1.12	8.91
	(9.33)	(0.00)	(0.02)	(0.87)	(6.89)	(6.20)
E.S.P	8.67 *	0	0.11 ***	4.19 ***	2.04	3.21
	(3.62)	(0.00)	(0.01)	(0.47)	(2.66)	(2.47)

*Notes: This table reports the regression results of $CSAD_t$ based on eq. (4.6), using the MD positive and negative tone values as further control variables in that equation. The data range is from 01/2008 to 10/2019. Numbers in the parentheses are standard errors. Standard errors are corrected for autocorrelation and heteroskedasticity using the Newey-West method [243]. Significance at 1%, 5%, 10%, 15% are denoted respectively by ***,**,*,*.*

ECB introductory statements tones are more pessimistic, we can speculate about the fact that herding behaviour is more likely to occur because of (among other features) *this* investors' characteristic.

4.4 Concluding remarks and insights for future research

This thesis investigated whether ECB communication can improve our understanding of the European stock market during Govern Council meeting days. Precisely, it analyses: (i) the effects that a generic press conference sentiment about specific topics can have on contemporaneous value of STOXX50 realizations; (ii) how herding behaviour across different European financial sectors can be detected by means of an augmented Chiang and Zeng [66] regression, that takes into account ECB negative and positive tone.

Results show that ECB press conference tone can improve our understanding of *both* these kinds of phenomena. In particular, in the sample period that spans from the 01/2008 to 10/2019, the European stock market (as proxied from the STOXX50 realizations) seems to be affected from ECB opinions about corporate, financial stability and economic topics. In particular, ECB introductory statement tone about the aforesaid topics is *positively* associated with equity levels, indicating that more (less) positive tone about these arguments implies, on average, higher (lower) values for the STOXX50 on the same day of the ECB press conference release. Moreover, during the same sample period, it was found that investor in the Healthcare sector and the whole European market seem to herd around positive and negative tone of the ECB. We noted, in particular, that investor herd in the former sector around negative statements related to the economic outlook and positive tone about the financial stability of the system. In the second case (and this could be more problematic under a regulatory point of view, as we explained in section 3.2.2) the whole EU market (here proxied from an equally weighted portfolio of different EU sector benchmark indices) herd around positive tone about the positive Governing Council economic outlook.

The main contribution to the current literature that applies computational linguistics tools to analyse central bank communication is thus twofold. The first one is that the researcher should take into account not only the topical structure of a generic document regarding a Public Institution, but also its "historical" path and changes behind its type of communication. Specifically, my work is the first in applying a financial stability dictionary in order to analyse ECB tone about this kind of topic. Such hypothesis arises from the fact that on 2012 the ECB was officially empowered to pursue supervision of credit institutions in order to foster *financial stability* tasks at the European level, beside its role of monetary policy [92]. By not taking into account such a feature, a researcher attempting in analysing a possible relationship between the STOXX50 and ECB communication could lose relevant information that, instead, may be embedded into stock prices. Second, I introduced an augmented regression with respect to the one proposed from Chiang and Zeng [66]. As we said in chapter 3, among other things, investor could be affected by *words* [148]. It is then important to consider also textual data in order to improve our understanding of herding phenomena in the financial markets, that could be "triggered" by relevant authorities such as the ECB. Again, Fig. 2.1b was clearly an empirical example of such a pattern.

At the same time, there are several avenues for extending the analysis, both toward the asset pricing and herding considerations. Given the high "flexibility" that such a text mining approach provides, in section 4.4.1 we will see at possible considerations about asset pricing, while in section 4.4.2 other ideas about how sentiment analysis can be used to detect herd behaviour will be proposed.

4.4.1 Considerations about ECB Sentiment Analysis

As we explained in section 4.3.1 different regressions were applied to the standardized values of the STOXX50, using the several sentiment time series as explanatory variables. However, different improvements could be applied in order to foster our understanding of how central bank tone can affect the stock market dynamics.

First, in each regression explained in eq. (4.5), *only* one variable was used to analyse the relationship between STOXX50 values and ECB tone. One could augment the same regression to take into account for other control variables, both macroeconomic [261], and financial ones [182], [288].

Second, it would be interesting to replicate a similar text mining procedure for the FED statements, analysing whether its impact on the stock market shows a *spill over effect* (see section 2.2.3 and 3.2.1). In other words, as we showed in Fig. 2.9 this effect has been found in the literature only by means of an indirect approach (see section 2.2.1). However, such a text mining analysis could provide other valuable information in order to understand the European stock market dynamics.

Third, as stated from Carrannante et al. [59]: "*Several authors have shown better results in forecasting economic variables by considering the sentiment values in their models.* ***Few studies have focused on the identification of the causes which explain opinions and beliefs***". It would then be interesting to understand the quantitative variables that could drive such a type of ECB communication, thus "inverting" the logic followed in eq. (4.5). Notably, this would allow us to comprehend better the several patterns observed in Fig. 4.4 during time.

4.4.2 Consideration about Herding behaviour across EU sectors

Detecting herding behaviour in the financial market is one of the most challenging tasks to address in Economics and Finance. In this thesis, I followed the idea of Ren and Wu [275], who proposed a method to detect herd behaviour by means of a Sentiment Analysis approach. As they suggest, detecting herding in this way could improve our understanding of such a phenomenon given the fact that it is highly linked with *human psychology* [82] and *sociology* [255], [267].

This thesis just added to the literature another tool to detect herding in the European stock market, focussing (for the first time) on the direct impact of ECB communication. However, my eq. (4.6) can be implemented in several ways.

First, in order to analyse the direct impact of ECB communication on herding behaviour across European equity sectors, it would be interesting to use, as Ren and Wu [275] suggest, higher frequency data (like for instance data at hour or minute pace). This would allow us to split the effect of ECB communication on herding behaviour from other exogenous factors (or events) of that specific day.

Second, and to link the analysis of the possible FED spill over effects on the European market, eq. (4.6) could be extended to take into account positive and negative tone from its statements. As we explained in section 3.2.2, herding is one of the indicators for systemic risk potential [318]: it would then be valuable for both the FED, the ECB and other related Institutional bodies to monitor such a phenomenon to ensure the stability of the overall financial system.

Third, and maybe even more important, it would be interesting to discern between rational and irrational herding by looking at fundamentals data of a specific sector and comparing them to the results from eq. (4.6). This would ensure a more qualitative interpretation of such a pattern in financial markets. For instance, we could expect that more leveraged sectors could be negatively (positively) affected from negative (positive) tone about ECB financial stability issues. In turn, we could expect that investors (holding a large part of related companies in their portfolio) will herd around those specific days.

Appendix A

R-Code to remove special characters in Twitter data

The presence of a special character is problematic for computational tractability, in particular for text mining tasks. However, specific functions can be developed to deal with these situations and the one proposed here responds to this need[1]. In particular, the function `gsub()` allows us to select specific *regex* (see section 1.4.7) to remove them form textual (in this case Twitter) data.

```
Textprocessing <- function(x){
  gsub("http[[:alnum:]]*",'', x)
  gsub('http\\S+\\s*', '', x) # Remove URLs
  gsub('\\b+RT', '', x) # Remove RT
  gsub('#\\S+', '', x) # Remove Hashtags
  gsub('@\\S+', '', x) # Remove Mentions
  gsub('[[:cntrl:]]', '', x) # Remove Special characters
  gsub("\\d", '', x) # Remove other Special characters
  gsub('[[:punct:]]', '', x) # Remove Punctuations
  gsub("^[[:space:]]*","",x) # Remove leading whitespaces
  gsub("[[:space:]]*$","",x) # Remove trailing whitespaces
  gsub('␣+','␣',x) # Remove extra whitespaces
}
```

[1] https://gist.github.com/CateGitau/05e6ff80b2a3aaa58236067811cee44e

Appendix B

Stopwords of the ECB press conferences

As explained in subsection 1.5.3, stopwords could lower the text mining quality and increase the computational efforts of our algorithm [339].

This is true also for a standardized communication as the one by the ECB. Hence, as we can see from the code below, the following list of stopwords were created, both for the introductory statement (`words.ecb`) and Q&A section (`words.qa`):

```
words.ecb <- c("Ladies", "gentlemen", "Vice", "President",
               "welcome", "Governing", "Council", "ECB",
               "detail,", "disposal", "questions", "almunia on"
                  )
```

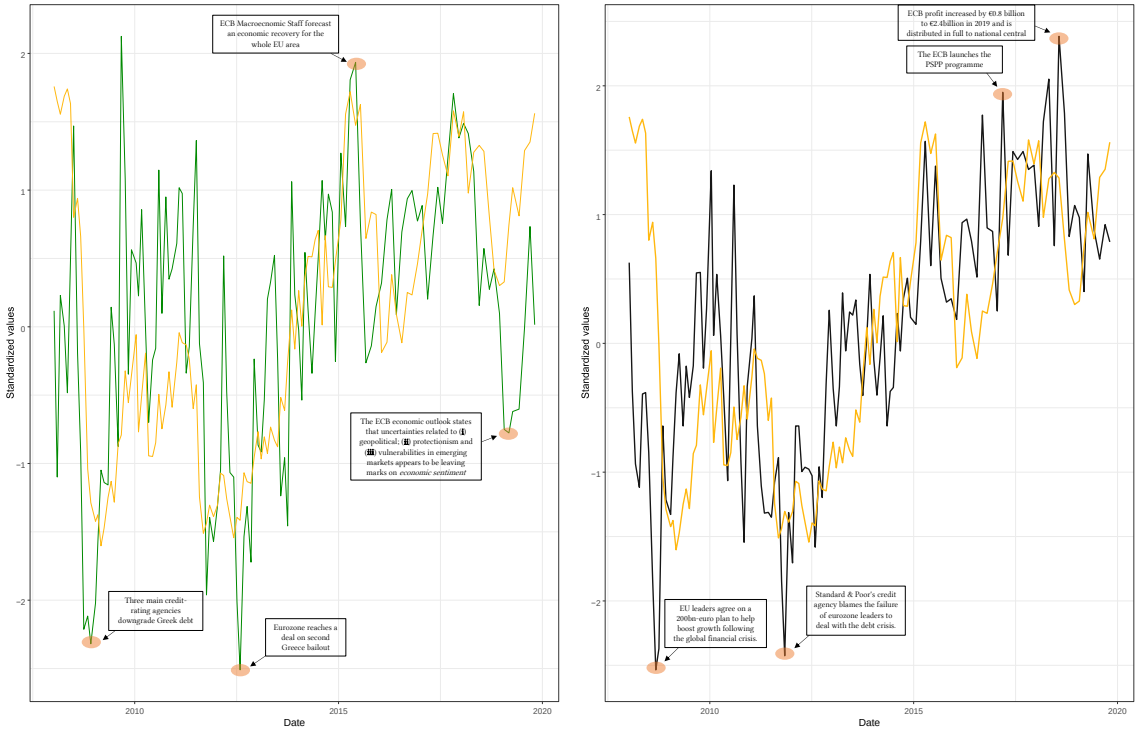
```
words.qa <- c("question", "gentlemen", "Vice", "President",
              "welcome", "Governing", "Council", "ECB",
              "detail,", "disposal", "questions", "it s",
              "that s", "draghi", "trichet", "van",
              "governing", "council", "ecb", "stupid",
              "setters", "beno t", "rompuy", "question the",
              "they re", "that draghi", "question draghi",
              "council the")
```

Notably, some of these stopwords have been selected after the tokenization [225] process applied by means of the `unnest_tokens()` function. Their "strange" semantic meaning could be due to the stemming algorithm (see section 1.5.3) embodied in the aforesaid function [338].

Appendix C

Sentiment series for the E.C. and M.D. lexicons and EU events

Fig. C.1 exhibits the EC and MD series and the major events in the EU and ECB history. As explained in section 4.2.2, albeit both series do not always show a (quantitative) significant relationship with the STOXX50, they positive and negative spikes allow us to detect important events in the EU history or in the European economic wealth state.



(a) The EC sentiment time series and the STOXX50 **(b)** The MD sentiment time series and the STOXX50

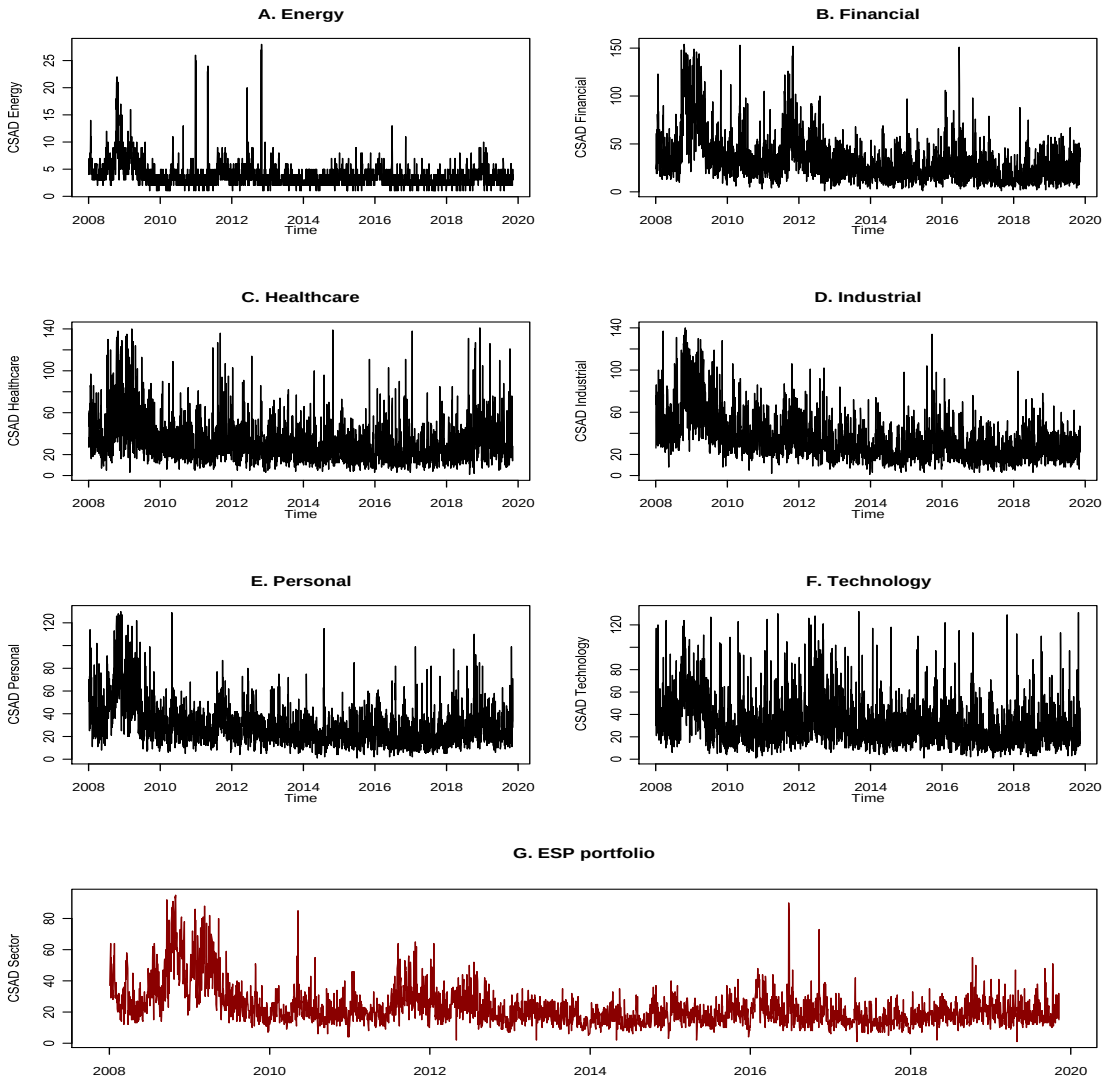
Figure C.1: A representation of a general ECB press release

Appendix D

Graphical representation of CSAD values for each EU sector

Fig. D.1 shows $CSAD_t$ values for several European financial sectors and for the ESP portfolio. Their examination allow us to understand the statistical values in table 4.4.

Figure D.1: $CSAD_t$ values across different European financial sectors



Bibliography

[1] Shigeo Abe and Takuya Inoue. "Fuzzy support vector machines for multiclass problems". In: *ESANN*. 2002, pp. 113–118.

[2] Serge Abiteboul. "Querying semi-structured data". In: *International Conference on Database Theory*. Springer. 1997, pp. 1–18.

[3] Viral V Acharya and Tanju Yorulmazer. "Information contagion and bank herding". In: *Journal of Money, Credit and Banking* 40.1 (2008), pp. 215–231.

[4] George A Akerlof and Robert J Shiller. *Animal spirits: How human psychology drives the economy, and why it matters for global capitalism*. Ed. by Princeton University Press. Princeton University Press, 2010.

[5] Md Akhtaruzzaman, Sabri Boubaker and Ahmet Sensoy. "Financial contagion during COVID-19 crisis". In: *Finance Research Letters* (2020), p. 101604.

[6] Andres Algaba et al. "Econometrics meets sentiment: An overview of methodology and applications". In: *Journal of Economic Surveys, Forthcoming* (2020).

[7] Ethem Alpaydin. *Introduction to Machine Learning*. Ed. by MIT Press. MIT Press, 2020.

[8] Diego Amaya and Jean-Yves Filbien. "The similarity of ECB's communication". In: *Finance Research Letters* 13 (2015), pp. 234–242.

[9] Werner Antweiler and Murray Z Frank. "Is all that talk just noise? The information content of internet stock message boards". In: *The Journal of Finance* 59.3 (2004), pp. 1259–1294.

[10] Mikael Apel and Marianna Grimaldi. "The information content of central bank minutes". In: *Riksbank Research Paper Series* 92 (2012).

[11] Jose Apesteguia, Simon Weidenholzer and Jörg Oechssler. "Copy trading". In: *Management Science* (2019).

[12] David Ardia et al. *The R package sentometrics to compute, aggregate and predict with textual sentiment*. Tech. rep. SSRN. 3067734. Working paper, 2017.

[13] Taylor B Arnold. "kerasR: R interface to the keras deep learning library". In: *Journal of Open Source Software* 2.14 (2017), p. 296.

[14] Masahiro Ashiya and Takero Doi. "Herd behavior of Japanese economists". In: *Journal of Economic Behavior & Organization* 46.3 (2001), pp. 343–346.

[15] Christopher N Avery and Judith A Chevalier. "Herding over the career". In: *Economics Letters* 63.3 (1999), pp. 327–333.

[16] Ian Ayres and Joshua Mitts. "Anti-Herding Regulation". In: *Harvard Business Law Review* 5 (2015), p. 1.

[17] Michelle Baddeley. "Herding, social influence and economic decision-making: socio-psychological and neuroscientific analyses". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.1538 (2010), pp. 281–290.

[18] Michelle Baddeley. "Herding, social influence and expert opinion". In: *Journal of Economic Methodology* 20.1 (2013), pp. 35–44.

[19] Michelle Baddeley et al. *Herding in financial behaviour: A behavioural and neuroeconomic analysis of individual differences*. Tech. rep. Faculty of Economics, University of Cambridge, 2012.

[20] Mark Bagnoli, Messod D Beneish and Susan G Watts. "Whisper forecasts of quarterly earnings per share". In: *Journal of Accounting and Economics* 28.1 (1999), pp. 27–50.

[21] Richard T Baillie. "Long memory processes and fractional integration in econometrics". In: *Journal of Econometrics* 73.1 (1996), pp. 5–59.

[22] Richard T Baillie. "Predictions from ARMAX models". In: *Journal of Econometrics* 12.3 (1980), pp. 365–374.

[23] Scott R Baker, Nicholas Bloom and Steven J Davis. "Measuring economic policy uncertainty". In: *The Quarterly Journal of Economics* 131.4 (2016), pp. 1593–1636.

[24] Scott R Baker et al. *Covid-induced economic uncertainty*. Tech. rep. National Bureau of Economic Research, 2020.

[25] Gökhan BakIr et al. *Predicting Structured Data*. Ed. by MIT Press. Neural Information ProcessingNeural Information Processing. MIT Press, 2007.

[26] Tanmay Bansal. "Behavioral Finance and COVID-19: Cognitive Errors that Determine the Financial Future". In: *Available at SSRN 3595749* (2020).

[27] Te Bao, Cees Diks and Hao Li. "A generalized CAPM model with asymmetric power distributed errors with an application to portfolio construction". In: *Economic Modelling* 68 (2018), pp. 611–621.

[28] Nicholas Barberis and Andrei Shleifer. "Style investing". In: *Journal of Financial Economics* 68.2 (2003), pp. 161–199.

[29] Michela Becchi and Patrick Crowley. "Efficient regular expression evaluation: theory to practice". In: *Proceedings of the 4th ACM/IEEE Symposium on Architectures for Networking and Communications Systems*. 2008, pp. 50–59.

[30] Stelios Bekiros et al. "Herding behavior, market sentiment and volatility: Will the bubble resume?" In: *The North American Journal of Economics and Finance* 42 (2017), pp. 107–131.

[31] Aymen Belgacem and Amine Lahiani. "Herding behavior around US macroeconomic announcements". In: *Journal of Applied Business Research (JABR)* 29.5 (2013), pp. 1401–1410.

[32] Hamza Bennani et al. "Does central bank communication signal future monetary policy in a (post)-crisis era? The case of the ECB". In: *Journal of International Money and Finance* (2020), p. 102167.

[33] Ahmed BenSaida, Mouna Jlassi and Houda Litimi. "Volume–herding interaction in the American market". In: *American Journal of Finance and Accounting* 4.1 (2015), pp. 50–69.

[34] Elisabeth SC Berger, Matthias Wenzel and Veit Wohlgemuth. "Imitation-related performance outcomes in social trading: A configurational approach". In: *Journal of Business Research* 89 (2018), pp. 322–327.

[35] Dan Bernhardt, Murillo Campello and Edward Kutsoati. "Who herds?" In: *Journal of Financial Economics* 80.3 (2006), pp. 657–675.

[36] David Bholat et al. *Sending firm messages: text mining letters from PRA supervisors to banks and building societies they regulate*. Tech. rep. Bank of England, 2017.

[37] David Bholat et al. *Text mining for central banks*. Tech. rep. Centre for Central Banking Studies, Bank of England, 2015.

[38] Fischer Black. "Capital market equilibrium with restricted borrowing". In: *The Journal of Business* 45.3 (1972), pp. 444–455.

[39] Lamont Black et al. "The systemic risk of European banks during the financial and sovereign debt crises". In: *Journal of Banking & Finance* 63 (2016), pp. 107–125.

[40] Natividad Blasco, Pilar Corredor and Sandra Ferreruela. "Detecting intentional herding: what lies beneath intraday data in the Spanish stock market". In: *Journal of the Operational Research Society* 62.6 (2011), pp. 1056–1066.

[41] Natividad Blasco, Pilar Corredor and Sandra Ferreruela. "Does herding affect volatility? Implications for the Spanish stock market". In: *Quantitative Finance* 12.2 (2012), pp. 311–327.

[42] David M Blei, Andrew Y Ng and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of Machine Learning Research* 3.Jan (2003), pp. 993–1022.

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

[44] Stephen P Borgatti and Daniel S Halgin. "On network theory". In: *Organization Science* 22.5 (2011), pp. 1168–1181.

[45] Claudio EV Borio and Boris Hofmann. *Is monetary policy less effective when interest rates are persistently low?* Tech. rep. Bank for International Settlements, 2017, p. 59.

[46] Benjamin Born, Michael Ehrmann and Marcel Fratzscher. "Central bank communication on financial stability". In: *The Economic Journal* 124.577 (2014), pp. 701–734.

[47] John F Boschen and Leonard O Mills. "The relation between narrative and money market indicators of monetary policy". In: *Economic Inquiry* 33.1 (1995), pp. 24–44.

[48] Margaret M Bradley and Peter J Lang. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Tech. rep. Technical report C-1, the Center for research in psychophysiology, 1999.

[49] Max Bramer. "An information-theoretic approach to the pre-pruning of classification rules". In: *International Conference on Intelligent Information Processing*. Springer. 2002, pp. 201–212.

[50] Max Bramer. *Principles of data mining*. Vol. 180. Springer, 2007.

[51] Claus Brand, Daniel Buncic and Jarkko Turunen. "The impact of ECB monetary policy decisions and communication on the yield curve". In: *Journal of the European Economic Association* 8.6 (2010), pp. 1266–1298.

[52] Eike Christain Brechmann and Claudia Czado. "Risk management with high-dimensional vine copulas: An analysis of the Euro Stoxx 50". In: *Statistics and Risk Modeling* 30.4 (2013), pp. 307–342.

[53] Markus K Brunnermeier and Martin Oehmke. "Bubbles, financial crises, and systemic risk". In: *Handbook of the Economics of Finance*. Vol. 2. Elsevier, 2013, pp. 1221–1288.

[54] Francesca Brusa, Pavel Savor and Mungo Wilson. "One central bank to rule them all". In: *Review of Finance* 24.2 (2020), pp. 263–304.

[55] Leland Bybee et al. *The structure of economic news*. Tech. rep. National Bureau of Economic Research, 2020.

[56] Esin Cakan and Aram Balagyozyan. "Sectoral herding: Evidence from an emerging market". In: *Journal of Accounting and Finance* 16.4 (2016).

[57] Esin Cakan et al. "Economic Policy Uncertainty and Herding Behavior: Evidence from the South African Housing Market". In: *Advances in Decision Sciences* 23.1 (2019), pp. 1–25.

[58] San Cannon et al. "Sentiment of the FOMC: Unscripted". In: *Economic Review-Federal Reserve Bank of Kansas City* 5 (2015).

[59] M Carrannante et al. "Temporal sentiment analysis with distributed lag models". In: *Smart Statistics for Smart Applications. SIS 2019*. Pearson. 2019, pp. 149–156.

[60] Rich Caruana and Alexandru Niculescu-Mizil. "An empirical comparison of supervised learning algorithms". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 161–168.

[61] Carlos Carvalho, Fernando Cordeiro and Juliana Vargas. "Just words?: A quantitative analysis of the communication of the Central Bank of Brazil". In: *Revista Brasileira de Economia* 67.4 (2013), pp. 443–455.

[62] Fabio Castiglionesi. "Financial contagion and the role of the central bank". In: *Journal of Banking & Finance* 31.1 (2007), pp. 81–101.

[63] Eric C Chang, Joseph W Cheng and Ajay Khorana. "An examination of herd behavior in equity markets: An international perspective". In: *Journal of Banking & Finance* 24.10 (2000), pp. 1651–1679.

[64] Irene Cherono, Tobias Olweny and Tabitha Nasieku. "Investor Behavior Biases and Stock Market Reaction in Kenya". In: *Journal of Applied Finance and Banking* 9.1 (2019), pp. 147–180.

[65] Yin-Wong Cheung and Kon S Lai. "Lag order and critical values of the augmented Dickey–Fuller test". In: *Journal of Business & Economic Statistics* 13.3 (1995), pp. 277–280.

[66] Thomas C Chiang and Dazhi Zheng. "An empirical analysis of herd behavior in global stock markets". In: *Journal of Banking & Finance* 34.8 (2010), pp. 1911–1921.

[67] Robert S Chirinko and Christopher Curran. *Greenspan Shrugs: Central Bank Communication, Formal Pronouncements and Bond Market Volatility*. Tech. rep. CESifo Working Paper, 2013, p. 149.

[68] C Chithra and E Ramaraj. "Heuristic sentence boundary detection and classification". In: *International Journal on Emerging Technologies* 7.2 (2016), pp. 199–206.

[69] Hyuk Choe, Bong-Chan Kho and Rene M Stulz. "Do foreign investors destabilize stock markets? The Korean experience in 1997". In: *Journal of Financial Economics* 54.2 (1999), pp. 227–264.

[70] Jay Pil Choi. "Herd behavior, the" penguin effect," and the suppression of informational diffusion: an analysis of informational externalities and payoff interdependency". In: *The Rand Journal of Economics* (1997), pp. 407–425.

[71] Nicole Choi and Richard W Sias. "Institutional industry herding". In: *Journal of Financial Economics* 94.3 (2009), pp. 469–491.

[72] William G Christie and Roger D Huang. "Following the pied piper: Do individual returns herd around the market?" In: *Financial Analysts Journal* 51.4 (1995), pp. 31–37.

[73] Günter Coenen et al. *Communication of monetary policy in unconventional times*. Tech. rep. ECB Working Paper, 2017.

[74] Meri Coleman and Ta Lin Liau. "A computer readability formula designed for machine scoring." In: *Journal of Applied Psychology* 60.2 (1975), p. 283.

[75] Robert Cooley, Bamshad Mobasher and Jaideep Srivastava. "Web mining: Information and pattern discovery on the world wide web". In: *Proceedings ninth IEEE international conference on tools with artificial intelligence*. IEEE. 1997, pp. 558–567.

[76] Ricardo Correa et al. *Sentiment in Central Banks' Financial Stability Reports*. Tech. rep. Board of Governors of the Federal Reserve System (US), 2017.

[77] Anacleto Correia, M Filomena Teodoro and Victor Lobo. "Statistical Methods for Word Association in Text Mining". In: *Recent Studies on Risk Analysis and Statistical Modeling*. Springer, 2018, pp. 375–384.

[78] Joshua D Coval and Tobias J Moskowitz. "Home bias at home: Local equity preference in domestic portfolios". In: *The Journal of Finance* 54.6 (1999), pp. 2045–2073.

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

[80] Alfred Cowles. "Can stock market forecasters forecast?" In: *Econometrica: Journal of the Econometric Society* (1933), pp. 309–324.

[81] Maintainer Gabor Csardi. "The igraph software package for complex network research". In: *InterJournal, complex systems* 1695.5 (2006), pp. 1–9.

[82] Viktoria Dalko. "Perception alignment hypothesis: causality of herding?" In: *Qualitative Research in Financial Markets* (2016).

[83] Mohammed Lawal Danrimi, Mazni Abdullah and Ervina Alfan. "Investors' herding practice: do IFRS and national economic culture matter?" In: *Managerial Finance* (2018).

[84] S. R. Das. *Text and Context : Language Analytics in Finance*. Ed. by Foundations and Trends® in Finance. Foundations and Trends® in Finance, 2014.

[85] Sanjiv R Das and Mike Y Chen. "Yahoo! for Amazon: Sentiment extraction from small talk on the web". In: *Management Science* 53.9 (2007), pp. 1375–1388.

[86] Nishant Dass, Massimo Massa and Rajdeep Patgiri. "Mutual funds and bubbles: The surprising role of contractual incentives". In: *The Review of Financial Studies* 21.1 (2008), pp. 51–99.

[87] Simon De Deyne and Gert Storms. "Word associations: Network and semantic properties". In: *Behavior Research Methods* 40.1 (2008), pp. 213–231.

[88] Gus De Franco et al. "Analyst report readability". In: *Contemporary Accounting Research* 32.1 (2015), pp. 76–104.

[89] Paul De Grauwe and Yuemei Ji. *Structural reforms and Monetary policies in a Behavioural Macroeconomic model*. Tech. rep. CEPR Discussion Papers, 2017.

[90] Joseph De la Vega. *Confusión de confusiones: diálogos curiosos entre un philosopho agudo, un mercader discreto, y un accionista erudito descriviendo el negocio de las acciones, su origen, su ethimologia, su realidad, su juego y su enredo*. Ed. by Editorial Maxtor Librería. 2009.

[91] Sandrine De Moerloose and Pierre Giot. "Style investing and momentum investing: A case study". In: *Journal of Asset Management* 12.6 (2011), pp. 407–417.

[92] Mattero De Poli. *Fundamentals of European banking law*. Ed. by CEDAM. First. Pubblicazioni della Facolta di Giurisprudenza dell'Universita di PadovaPubblicazioni della Facolta di Giurisprudenza dell'Universita di Padova. 2018.

[93] Matthew J Denny and Arthur Spirling. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It". In: *Political Analysis* 26.2 (2018), pp. 168–189.

[94] Andrea Devenow and Ivo Welch. "Rational herding in financial economics". In: *European Economic Review* 40.3-5 (1996), pp. 603–615.

[95] Hans Dewachter et al. "The intra-day impact of communication on euro-dollar volatility and jumps". In: *Journal of International Money and Finance* 43 (2014), pp. 131–154.

[96] Atanu Dey, Mamata Jenamani and Jitesh J Thakkar. "Senti-N-Gram: An n-gram lexicon for sentiment analysis". In: *Expert Systems with Applications* 103 (2018), pp. 92–105.

[97] Evgenia Dimitriadou et al. *The e1071 package*. 2006. URL: `https://cran.r-project.org/web/packages/e1071/index.html`.

[98] Annette J Dobson and Adrian G Barnett. *An introduction to generalized linear models*. Ed. by Chapman and Hall/CRC. CRC Press, 2018.

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

[100] John C Driscoll and Steinar Holden. "Behavioral Economics and Macroeconomic models". In: *Journal of Macroeconomics* 41 (2014), pp. 133–147.

[101] Michael Ehrmann and Marcel Fratzscher. "Explaining Monetary Policy in Press Conferences". In: *International Journal of Central Banking* 5.2 (2009), pp. 42–84.

[102] Robert F Engle et al. "Hedging climate change news". In: *The Review of Financial Studies* 33.3 (2020), pp. 1184–1216.

[103] A Erdenetsogt and V Kallinterakis. "Investors' herding in frontier markets: Evidence from Mongolia". In: *Handbook of Frontier Markets*. Elsevier, 2016, pp. 233–249.

[104] Eugene F Fama and Kenneth R French. "Industry costs of equity". In: *Journal of Financial Economics* 43.2 (1997), pp. 153–193.

[105] Federico Favaretto and Donato Masciandaro. "Behavioral economics and monetary policy". In: *BAFFI CAREFIN Centre Research Paper* 2015-1 (2015).

[106] Federico Favaretto and Donato Masciandaro. "Doves, hawks and pigeons: Behavioral monetary policy and interest rate inertia". In: *Journal of Financial Stability* 27 (2016), pp. 50–58.

[107] Federico Favaretto and Donato Masciandaro. "Populism, Group Thinking and Banking Policy". In: *BAFFI CAREFIN Centre Research Paper* 2020-133 (2020).

[108] Ingo Feinerer. *Introduction to the tm Package Text Mining in R*. 2013.

[109] Ian Fellows et al. *Package 'wordcloud'*. 2018. URL: `https://cran.r-project.org/web/packages/wordcloud/index.html`.

[110] Eilis Ferran and Valia SG Babis. "The European Single Supervisory Mechanism". In: *Journal of Corporate Law Studies* 13.2 (2013), pp. 255–285.

[111] Stefan Feuerriegel, Nicolas Proellochs and Maintainer Stefan Feuerriegel. *Package 'SentimentAnalysis'*. 2018. URL: `https://cran.r-project.org/web/packages/SentimentAnalysis/vignettes/SentimentAnalysis.html`.

[112] Andy Field. *Discovering statistics using SPSS*. Ed. by SAGE. Vol. 264. Sage Publications Chennai, India, 2009, p. 315.

[113] William Forbes. *Behavioural Finance*. Ed. by Wiley. John Wiley & Sons, 2009.

[114] W Nelson Francis and Henry Kucera. "Brown corpus manual". In: *Letters to the Editor* 5.2 (1979), p. 7.

[115] Emilios C Galariotis, Styliani-Iris Krokida and Spyros I Spyrou. "Bond market investor herding: Evidence from the European financial crisis". In: *International Review of Financial Analysis* 48 (2016), pp. 367–375.

[116] Priyank Gandhi, Tim Loughran and Bill McDonald. "Using annual report sentiment as a proxy for financial distress in US banks". In: *Journal of Behavioral Finance* 20.4 (2019), pp. 424–436.

[117] R Gaston Gelos and Shang-Jin Wei. "Transparency and international portfolio holdings". In: *The Journal of Finance* 60.6 (2005), pp. 2987–3020.

[118] Matthew Gentzkow, Bryan Kelly and Matt Taddy. "Text as data". In: *Journal of Economic Literature* 57.3 (2019), pp. 535–74.

[119] Erik Gerding. *Law, Bubbles, and Financial Regulation*. Ed. by Routledge. Routledge, 2013.

[120] Zoubin Ghahramani. "Unsupervised Learning". In: *Summer School on Machine Learning*. Springer. 2003, pp. 72–112.

[121] Diman Ghazi, Diana Inkpen and Stan Szpakowicz. "Hierarchical versus flat classification of emotions in text". In: *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. Association for Computational Linguistics. 2010, pp. 140–146.

[122] Eric Ghysels and Massimiliano Marcellino. *Applied economic forecasting using time series methods*. Ed. by OUP USA. Oxford University Press, 2018.

[123] Wael H Gomaa, Aly A Fahmy et al. "A survey of text similarity approaches". In: *International Journal of Computer Applications* 68.13 (2013), pp. 13–18.

[124] Pu Gong and Jun Dai. "Monetary policy, exchange rate fluctuation, and herding behavior in the stock market". In: *Journal of Business Research* 76 (2017), pp. 34–43.

[125] Bryan Goodrich, Dason Kurkiewicz and Tyler Rinker. *Package 'qdap'*. 2018. URL: `https://cran.r-project.org/web/packages/qdap/index.html`.

[126] Clive WJ Granger. "Spurious regressions in econometrics". In: *Journal of Econometrics* 2.2 (1974), pp. 111–120.

[127] Mark Grinblatt, Sheridan Titman and Russ Wermers. "Momentum investment strategies, portfolio performance, and herding: A study of mutual fund behavior". In: *The American Economic Review* (1995), pp. 1088–1105.

[128] David Gruen, Michael Plumb and Andrew Stone. "How should monetary policy respond to asset-price bubbles?" In: *International Journal of Central Banking* (2005).

[129] Robert Gunning et al. *Technique of Clear Writing*. Ed. by McGraw-Hill. 1952.

[130] Dinesh Gupta. "Herding Behavior in Financial Market: Critical Literature Review". In: *Asian Journal of Research in Banking and Finance* 8.6 (2018), pp. 60–72.

[131] Refet S Gürkaynak, Brian P Sack and Eric T Swanson. "Do actions speak louder than words? The response of asset prices to monetary policy actions and statements". In: *International Journal of Central Banking* (2004), pp. 55–93.

[132] M Haigh, N Boyd and Bahattin Buyuksahin. "Herding amongst hedge funds in futures markets". In: *Commodity Futures Trading Commission* (2006).

[133] Peter Haiss. "Bank herding and incentive systems as catalysts for the financial crisis". In: *IUP Journal of Behavioral Finance* 7.1/2 (2010), p. 30.

[134] Peter R Haiss. "Banks, herding and regulation: A review and synthesis". In: *Paper for presentation at the workshop on Informational Herding Behavior, Copenhagen*. 2005.

[135] Haifa Hammami and Younes Boujelbene. "Investor Herding Behavior and Its Effect on Stock Market Boom-Bust Cycles." In: *IUP Journal of Applied Finance* 21.1 (2015).

[136] Stephen Hansen and Michael McMahon. "Shocking language: Understanding the macroeconomic effects of central bank communication". In: *Journal of International Economics* 99 (2016), S114–S133.

[137] John Harrison and Maintainer John Harrison. *Package 'RSelenium'*. 2020. URL: `https://cran.r-project.org/web/packages/RSelenium/vignettes/basics.html`.

[138] James E Hartley and James E Hartley. *The Representative Agent in Macroeconomics*. Ed. by Routledge. First. 2002.

[139] Philipp Hartmann and Frank Smets. *The first twenty years of the European Central Bank: monetary policy*. Tech. rep. European Central Bank, 2018.

[140] Chikio Hayashi. "What is Data Science? Fundamental concepts and a heuristic example". In: *Data science, classification, and related methods*. Springer, 1998, pp. 40–51.

[141] Zhiguo He and Arvind Krishnamurthy. "A macroeconomic framework for quantifying systemic risk". In: *American Economic Journal: Macroeconomics* 11.4 (2019), pp. 1–37.

[142] Marti A Hearst. "Untangling text data mining". In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics. 1999, pp. 3–10.

[143] Ahmed Hefnaoui et al. "Analysis of herding behavior in Moroccan stock market". In: *Journal of Economics and Behavioral Studies* 11.1 (J) (2019), pp. 181–190.

[144] Friedrich Heinemann and Katrin Ullrich. "Does it pay to watch central bankers' lips? The information content of ECB wording". In: *Swiss Journal of Economics and Statistics* 143.2 (2007), pp. 155–185.

[145] Ottar Hellevik. "Linear versus logistic regression when the dependent variable is a dichotomy". In: *Quality & Quantity* 43.1 (2009), pp. 59–74.

[146] Scott Hendry and Alison Madeley. *Text mining and the information content of Bank of Canada communications*. Tech. rep. Bank of Canada Working Paper, 2010.

[147] Ralf Herbrich, Thore Graepel, Klaus Obermayer et al. *Regression models for ordinal data: A Machine Learning approach*. Citeseer, 1999.

[148] David Hirshleifer and Siew Hong Teoh. "Herd behaviour and cascading in capital markets: A review and synthesis". In: *European Financial Management* 9.1 (2003), pp. 25–66.

[149] Bart Hobijn, Philip Hans Franses and Marius Ooms. "Generalizations of the KPSS-test for stationarity". In: *Statistica Neerlandica* 58.4 (2004), pp. 483–502.

[150] Mark Hodnett and Joshua F Wiley. *R Deep Learning Essentials: A step-by-step guide to building deep learning models using TensorFlow, Keras, and MXNet*. Packt Publishing Ltd, 2018.

[151] Rani Hoitash, Udi Hoitash and Landi Morris. "eXtensible Business Reporting Language: A Review and Directions for Future Research". In: *Available at SSRN* (2020).

[152] Phil Holmes, Vasileios Kallinterakis and MP Leite Ferreira. "Herding in a concentrated market: a question of intent". In: *European Financial Management* 19.3 (2013), pp. 497–520.

[153] Cars H Hommes. *Behavioral and experimental macroeconomics and policy analysis: A complex systems approach*. Tech. rep. ECB Working Paper, 2018.

[154] Harrison Hong, Jeffrey D Kubik and Amit Solomon. "Security analysts' career concerns and herding of earnings forecasts". In: *The Rand Journal of economics* (2000), pp. 121–144.

[155] Harrison Hong, Jeffrey D Kubik and Jeremy C Stein. "Thy neighbor's portfolio: Word-of-mouth effects in the holdings and trades of money managers". In: *The Journal of Finance* 60.6 (2005), pp. 2801–2824.

[156] Shu-Fan Hsieh. "Individual and institutional herding and the impact on stock returns: Evidence from Taiwan stock market". In: *International Review of Financial Analysis* 29 (2013), pp. 175–188.

[157] Chun-Nan Hsu and Chien-Chi Chang. "Finite-state transducers for semi-structured text mining". In: *Proceedings of IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications*. Menlo Park, CA, USA IJCAI Co. 1999, pp. 38–49.

[158] Anna Huang. "Similarity measures for text document clustering". In: *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*. Vol. 4. 2008, pp. 9–56.

[159] Yifen Huang and Tom M Mitchell. "Text clustering with extended user feedback". In: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 2006, pp. 413–420.

[160] M Humayun Kabir. "Did investors herd during the financial crisis? Evidence from the US financial industry". In: *International Review of Finance* 18.1 (2018), pp. 59–90.

[161] Weifeng Hung, Chia-Chi Lu and Cheng F Lee. "Mutual fund herding its impact on stock returns: Evidence from the Taiwan stock market". In: *Pacific-Basin Finance Journal* 18.5 (2010), pp. 477–493.

[162] Pam Hurley. *Readability*. Tech. rep. Hurley Write, 2016.

[163] Soosung Hwang and Mark Salmon. "Underconfidence, Pessimism and the Low-Beta Anomaly". In: *Available at SSRN 299919* (2017).

[164] Edgars Rihards Indārs, Aliaksei Savin and Ágnes Lublóy. "Herding behaviour in an emerging market: Evidence from the Moscow Exchange". In: *Emerging Markets Review* 38 (2019), pp. 468–487.

[165] Mohit Iyyer et al. "A neural network for factoid question answering over paragraphs". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 633–644.

[166] Brandon Janos Barberis Larry Cao et al. *Fintech 2018: The ASIA Pacific Edition*. Tech. rep. CFA Institute, 2018.

[167] David-Jan Jansen and Jakob de Haan. *The importance of being vigilant: has ECB communication influenced Euro area inflation expectations?* Tech. rep. CESifo Working Paper, 2007.

[168] Narasimhan Jegadeesh and Di Wu. "Word power: A new approach for content analysis". In: *Journal of Financial Economics* 110.3 (2013), pp. 712–729.

[169] Narasimhan Jegadeesh and Di Andrew Wu. "Deciphering Fedspeak: The information content of FOMC meetings". In: *Available at SSRN 2939937* (2017).

[170] Saif Ullah Jhandir and Muhammad Nadeem Hanif. "Herding Behavior Around Macroeconomic Announcements: An Evidence from Pakistan". In: *SZABIST's 20th National Research Conference, 10th May*. 2014.

[171] Yawen Jiao and Pengfei Ye. "Mutual fund herding in response to hedge fund herding and the impacts on stock prices". In: *Journal of Banking & Finance* 49 (2014), pp. 131–148.

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

[173] Matthew Jockers. *Package 'syuzhet'*. 2017. URL: `https://cran.%20r-project.%20org/web/packages/syuzhet`.

[174] Anders Johansen and Didier Sornette. "Financial" anti-bubbles": Log-periodicity in gold and Nikkei collapses". In: *International Journal of Modern Physics C (IJMPC)* 10.04 (1999), pp. 563–575.

[175] Vasileios Kallinterakis and Greg N Gregoriou. "Herd behaviour: A survey". In: *Aestimatio: the IEB International Journal of Finance* (2017).

[176] Koichiro Kamada and Ko Miura. *Confidence Erosion and Herding Behavior in Bond Markets: An Essay on Central Bank Communication Strategy*. Tech. rep. Bank of Japan, 2014.

[177] Nont Kanungsukkasem and Teerapong Leelanupab. "Financial Latent Dirichlet Allocation (FinLDA): Feature extraction in text and data mining for financial time series prediction". In: *IEEE Access* 7 (2019), pp. 71645–71664.

[178] Kohei Kawamura et al. *Strategic Central Bank Communication: Discourse and Game-Theoretic Analyses of the Bank of Japan's Monthly Report*. Tech. rep. University of Tokyo, Graduate School of Economics, 2016.

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

[180] John Maynard Keynes. *The General Theory of Employment, Interest, and Money*. Ed. by Stellar Classics (5 May 2016). Springer, 1936.

[181] Ashraf Khan. *A Behavioral Approach to Financial Supervision, Regulation, and Central Banking*. Tech. rep. International Monetary Fund, 2018.

[182] Justyna Klejdysz, Robin L Lumsdaine and Michel van der Wel. "Shifts in ECB communication: a text mining approach". MA thesis. Erasmus University Rotterdam, Aug. 2018.

[183] Antonina Kloptchenko et al. "Combining data and text mining techniques for analysing financial reports". In: *Intelligent Systems in Accounting, Finance & Management: International Journal* 12.1 (2004), pp. 29–41.

[184] Puput Tri Komalasari, Marwan Asri and Bowo Setiyono. "Bibliometric Analysis of Herding Behavior in Capital Market". In: *3rd Asia Pacific International Conference of Management and Business Science (AICMBS 2019)*. Atlantis Press. 2020, pp. 226–232.

[185] Sotiris B Kotsiantis, I Zaharakis and P Pintelas. "Supervised Machine Learning: A review of classification techniques". In: *Emerging artificial intelligence applications in computer engineering* 160 (2007), pp. 3–24.

[186] Styliani-Iris Krokida, Panagiota Makrychoriti and Spyros Spyrou. "Monetary policy and herd behavior: International evidence". In: *Journal of Economic Behavior & Organization* 170 (2020), pp. 386–417.

[187] Max Kuhn. *The caret package*. 2012. URL: `https://cran.r-project.org/web/packages/caret/vignettes/caret.html`.

[188] Ankit Kumar et al. "Ask me anything: Dynamic memory networks for Natural Language Processing". In: *International conference on Machine Learning*. 2016, pp. 1378–1387.

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

[190] Satish Kumar and Nisha Goyal. "Behavioural biases in investment decision making–a systematic literature review". In: *Qualitative Research in Financial Markets* 7.1 (2015), pp. 88–108.

[191] Josef Lakonishok, Andrei Shleifer and Robert W Vishny. "The impact of institutional trading on stock prices". In: *Journal of Financial Economics* 32.1 (1992), pp. 23–43.

[192] Duncan Temple Lang and Maintainer Duncan Temple Lang. *Package 'XML'*. 2013. URL: `https://cran.r-project.org/web/packages/XML/index.html`.

[193] Paulo Lao and Harminder Singh. "Herding behaviour in the Chinese and Indian stock markets". In: *Journal of Asian Economics* 22.6 (2011), pp. 495–506.

[194] Gary Baker Larry Cao Rhodri Preece. *AI Piooners in Investment Management*. Tech. rep. CFA Institute, 2019.

[195] Harold D Lasswell and J Zvi Namenwirth. *The Lasswell Value Dictionary (3 vols.). New Haven: Yale University*. Tech. rep. Mimeo, 1968.

[196] Chien-Chiang Lee, Mei-Ping Chen and Kuan-Mien Hsieh. "Industry herding and market states: evidence from Chinese stock markets". In: *Quantitative Finance* 13.7 (2013), pp. 1091–1113.

[197] Eunkyoung Lee and Byungtae Lee. "Herding behavior in online P2P lending: An empirical investigation". In: *Electronic Commerce Research and Applications* 11.5 (2012), pp. 495–503.

[198] David Leinweber and Jacob Sisk. "Event-driven trading and the "new news"". In: *The Journal of Portfolio Management* 38.1 (2011), pp. 110–124.

[199] Jim Lemon et al. "Plotrix: a package in the red light district of R". In: *R-news* 6.4 (2006), pp. 8–12.

[200] Zhuoqian Liang, Ding Pan and Yuan Deng. "Research on the Knowledge Association Reasoning of Financial Reports Based on a Graph Network". In: *Sustainability* 12.7 (2020), pp. 1–14.

[201] Aristidis Likas, Nikos Vlassis and Jakob J Verbeek. "The global k-means clustering algorithm". In: *Pattern Recognition* 36.2 (2003), pp. 451–461.

[202] Houda Litimi, Ahmed BenSaida and Omar Bouraoui. "Herding and excessive risk in the American stock market: A sectoral analysis". In: *Research in International Business and Finance* 38 (2016), pp. 6–21.

[203] Bing Liu and Kevin Chen-Chuan-Chang. "Special issue on Web content mining". In: *AcCM Sigkdd Explorations Newsletter* 6.2 (2004), pp. 1–4.

[204] Fan Liu. "Herd behavior in the insurance market: a survey". In: *International Journal of Economics and Finance* 7.11 (2015), pp. 154–162.

[205] Tim Loughran and Bill McDonald. "Measuring readability in financial disclosures". In: *The Journal of Finance* 69.4 (2014), pp. 1643–1671.

[206] Tim Loughran and Bill McDonald. "Textual analysis in accounting and finance: A survey". In: *Journal of Accounting Research* 54.4 (2016), pp. 1187–1230.

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

[208] Julie Beth Lovins. *Development of a stemming algorithm*. Tech. rep. 1-2. Massachusetts Institute of Tech Cambridge Electroinc system Lab, 1968, pp. 22–31.

[209] Hsin-Min Lu et al. "Financial text mining: Supporting decision making using web 2.0 content". In: *IEEE Intelligent Systems* 25.2 (2010), pp. 78–82.

[210] Pongsak Luangaram and Warapong Wongwachara. *More Than Words: A Textual Analysis of Monetary Policy Communication*. Tech. rep. Puey Ungphakorn Institute for Economic Research, 2017.

[211] David O Lucca and Francesco Trebbi. *Measuring central bank communication: an automated approach with application to FOMC statements*. Tech. rep. National Bureau of Economic Research, 2009.

[212] Hans Peter Luhn. "A business intelligence system". In: *IBM Journal of Research and Development* 2.4 (1958), pp. 314–319.

[213] Torben Lütje. "To be good or to be better: asset managers' attitudes towards herding". In: *Applied Financial Economics* 19.10 (2009), pp. 825–839.

[214] Ben Mabrouk Houda and Fakhfekh Mohamed. "Herding During Market Upturns and Downturns: International Evidence." In: *IUP Journal of Applied Finance* 19.2 (2013).

[215] Burton G Malkiel and Eugene F Fama. "Efficient capital markets: A review of theory and empirical work". In: *The Journal of Finance* 25.2 (1970), pp. 383–417.

[216] Anandadeep Mandal. "Empirical study of herd behavior: The national stock exchange, India". In: *International Journal of Financial Management* 1.3 (2011), pp. 1–11.

[217] Christopher Manning, Prabhakar Raghavan and Hinrich Schütze. "Introduction to information retrieval". In: *Natural Language Engineering* 16.1 (2010), pp. 100–103.

[218] Christopher D Manning, Christopher D Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. Ed. by MIT Press. MIT press, 1999.

[219] Christopher D Manning et al. "The Stanford CoreNLP natural language processing toolkit". In: *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 2014, pp. 55–60.

[220] Mika V Mäntylä, Daniel Graziotin and Miikka Kuutila. "The evolution of sentiment analysis - A review of research topics, venues, and top cited papers". In: *Computer Science Review* 27 (2018), pp. 16–32.

[221] Roberto Marchionatti. "On Keynes' Animal Spirits". In: *Kyklos* 52.3 (1999), pp. 415–439.

[222] Enrico Piero Marelli and Marcello Signorelli. *Politica economica: Le politiche nel nuovo scenario europeo e globale*. Ed. by Giappichelli (1 April 2015). G Giappichelli Editore, 2015.

[223] Mónica Marrero et al. "Named entity recognition: fallacies, challenges and opportunities". In: *Computer Standards & Interfaces* 35.5 (2013), pp. 482–489.

[224] Aakriti Mathur and Rajeswari Sengupta. *Analysing monetary policy statements of the Reserve Bank of India*. Tech. rep. Indira Gandhi Institute of Development Research, 2019.

[225] Ulf Mattsson and Yigal Rozenberg. *Tokenization in payment environments*. US Patent App. 13/761,009. 2013.

[226] DS Maylawati, H Aulawi and MA Ramdhani. "The concept of sequential pattern mining for text". In: *IOP Conference Series: Materials Science and Engineering*. Vol. 434. 1. IOP Publishing. 2018, p. 012042.

[227] Ryan McDonald and Fernando Pereira. "Identifying gene and protein mentions in text using conditional random fields". In: *BMC Bioinformatics* 6.S1 (2005), S6.

[228] Tony McEnery and Andrew Hardie. *Corpus linguistics: Method, theory and practice*. Ed. by Cambridge University Press. Cambridge University Press, 2011.

[229] Lukas Menkhoff, Ulrich Schmidt and Torsten Brozynski. "The impact of experience on risk taking, overconfidence, and herding of fund managers: Complementary survey evidence". In: *European Economic Review* 50.7 (2006), pp. 1753–1766.

[230] Charilaos Mertzanis and Noha Allam. "Political instability and herding behaviour: Evidence from Egypt's stock market". In: *Journal of Emerging Market Finance* 17.1 (2018), pp. 29–59.

[231] Petros Messis and Achilleas Zapranis. "Herding towards higher moment CAPM, contagion of herding and macroeconomic shocks: Evidence from five major developed markets". In: *Journal of Behavioral and Experimental Finance* 4 (2014), pp. 1–13.

[232] A Seddik Meziani. "Investing with Environmental, Social, and Governance Issuesin Mind: From the Back to the Fore of Style Investing". In: *The Journal of Investing* 23.3 (2014), pp. 115–124.

[233] Stefano Micossi, Alexandra D'Onofrio and Fabrizia Peirce. "Herd Behaviour in Asset Market Booms and Crashes: The Role of Monetary Policy". In: *Policy Insights* 97 (2019).

[234] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv* (2013).

[235] Gary Miner et al. *Practical text mining and statistical analysis for non-structured text data applications*. Ed. by Academic Press. 2012.

[236] Felix Ming et al. "Stock market prediction from WSJ: text mining via sparse matrix factorization". In: *2014 IEEE International Conference on Data Mining*. IEEE. 2014, pp. 430–439.

[237] Leela Mitra and Gautam Mitra. "Applications of news analytics in finance: A review". In: *The Handbook of News Analytics in Finance* 596.1 (2011).

[238] Javad Moradi and H Abbasi. "A test of investors' herding behavior in Tehran exchange". In: *Interdisciplinary Journal of Contemporary Research in Business* 3.10 (2012), pp. 686–702.

[239] Franco Moretti and Dominique Pestre. "Bankspeak: the language of World Bank reports". In: *New Left Review* 92.2 (2015), pp. 75–99.

[240] Venky Nagar and Jordan Schoenfeld. *Weather and Firm-Level Outcomes: New Evidence from a Linguistic Analysis*. Tech. rep. 3438428. Tuck School of Business Working Paper, 2019.

[241] Arman Khadjeh Nassirtoussi et al. "Text mining for market prediction: A systematic review". In: *Expert Systems with Applications* 41.16 (2014), pp. 7653–7670.

[242] Andrew A Neath and Joseph E Cavanaugh. "The Bayesian information criterion: background, derivation, and applications". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 4.2 (2012), pp. 199–203.

[243] Whitney K Newey and Kenneth D West. "Hypothesis testing with efficient method of moments estimation". In: *International Economic Review* (1987), pp. 777–787.

[244] Trong Dung Nguyen, Tu Bao Ho and Hiroshi Shimodaira. "A scalable algorithm for rule post-pruning of large decision trees". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2001, pp. 467–476.

[245] Sergey I Nikolenko, Sergei Koltcov and Olessia Koltsova. "Topic modelling for qualitative studies". In: *Journal of Information Science* 43.1 (2017), pp. 88–102.

[246] John R Nofsinger and Richard W Sias. "Herding and feedback trading by institutional and individual investors". In: *The Journal of Finance* 54.6 (1999), pp. 2263–2295.

[247] Clemens Nopp and Allan Hanbury. "Detecting risks in the banking system by sentiment analysis". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 591–600.

[248] Löfving Olof. "Sentiment Analysis of Equity Analyst Research Reports using Convolutional Neural Networks". MA thesis. Uppsala University Publications, 2019.

[249] Moatemri Ouarda, Abdelfatteh El Bouri and Olivero Bernard. "Herding behavior under markets condition: Empirical evidence on the European financial markets". In: *International Journal of Economics and Financial Issues* 3.1 (2013), p. 214.

[250] Krishna G Palepu, Paul M Healy and Erik Peek. *Business Analysis and Valuation*. Ed. by Cengage Learning EMEA. Fifth. 2019.

[251] Wei Pan. "Akaike's information criterion in generalized estimating equations". In: *Biometrics* 57.1 (2001), pp. 120–125.

[252] Bo Pang, Lillian Lee et al. "Opinion mining and sentiment analysis". In: *Foundations and Trends® in Information Retrieval* 2.1–2 (2008), pp. 1–135.

[253] Thomas Papadopoulos. "European System of Financial Supervision". In: *Max Planck Encyclopedia of Public International Law, Oxford University Press (2014)* (2015).

[254] Ki Young Park, Youngjoon Lee and Soohyon Kim. *Deciphering Monetary Policy Board Minutes through Text Mining Approach: The Case of Korea*. Tech. rep. Economic Research Institute, Bank of Korea, 2019.

[255] Wayne D Parker and Robert R Prechter. "Herding: An Interdisciplinary Integrative Review from a Socionomic Perspective". In: *in Kokinov, Boicho, Ed., Advances in Cognitive Economics: Proceedings of the International Conference on Cognitive Economics, Bulgaria: NBU Press (New Bulgarian University.* Citeseer. 2005.

[256] Duo Pei and Miklos A Vasarhelyi. "Big data and algorithmic trading against periodic and tangible asset reporting: The need for U-XBRL". In: *International Journal of Accounting Information Systems* (2020), p. 100453.

[257] Mirjana Pejić Bach et al. "Text mining for big data analysis in financial sector: A literature review". In: *Sustainability* 11.5 (2019), p. 1277.

[258] Francisco Pereira, Tom Mitchell and Matthew Botvinick. "Machine learning classifiers and fMRI: a tutorial overview". In: *Neuroimage* 45.1 (2009), S199–S209.

[259] Peter CB Phillips and Sam Ouliaris. "Asymptotic properties of residual based tests for cointegration". In: *Econometrica: Journal of the Econometric Society* (1990), pp. 165–193.

[260] Peter CB Phillips and Pierre Perron. "Testing for a unit root in time series regression". In: *Biometrika* 75.2 (1988), pp. 335–346.

[261] Matthieu Picault and Thomas Renault. "Words are not all created equal: A new measure of ECB communication". In: *Journal of International Money and Finance* 79 (2017), pp. 136–156.

[262] Christian Pierdzioch, Jan-Christoph Rülke and Georg Stadtmann. "Forecasting metal prices: Do forecasters herd?" In: *Journal of Banking & Finance* 37.1 (2013), pp. 150–158.

[263] Joël Plisson, Nada Lavrac, Dunja Mladenic et al. "A rule based approach to word lemmatization". In: *Proceedings of IS04*. Vol. 3. 2004, pp. 83–86.

[264] Brett Powley, Robert Dale et al. "Evidence-Based Information Extraction for High Accuracy Citation and Author Name Identification." In: *RIAO International Conference on Large-Scale Semantic Access to Content*. Vol. 7. Citeseer. 2007, pp. 618–632.

[265] Henriette Prast. *Herding and financial panics: a role for cognitive psychology?* Tech. rep. Netherlands Central Bank, Research Department, 2000.

[266] Jaya M Prosad, Sujata Kapoor and Jhumur Sengupta. "An examination of herd behavior: An empirical evidence from Indian equity market". In: *International Journal of Trade, Economics and Finance* 3.2 (2012), p. 154.

[267] Ramsey M Raafat, Nick Chater and Chris Frith. "Herding in humans". In: *Trends in Cognitive Sciences* 13.10 (2009), pp. 420–428.

[268] Claudio Raddatz and Sergio Schmukler. *Deconstructing herding: evidence from pension fund investment behavior*. Tech. rep. The World Bank, 2011.

[269] Faisal Rahutomo, Teruaki Kitasuka and Masayoshi Aritsugi. "Semantic cosine similarity". In: *The 7th International Student Conference on Advanced Science and Technology ICAST*. Vol. 4. 1. 2012.

[270] Raghu Ramakrishnan and Johannes Gehrke. *Database management systems*. Ed. by McGraw Hill. McGraw Hill, 2000.

[271] Angelo Ranaldo and Enzo Rossi. "The reaction of asset markets to Swiss National Bank communication". In: *Journal of International Money and Finance* 29.3 (2010), pp. 486–503.

[272] Tushar Rao and Saket Srivastava. "Analyzing Stock Market Movements Using Twitter Sentiment Analysis". In: *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. 2012, pp. 119–123.

[273] Jonathon Read et al. "Sentence boundary detection: A long solved problem?" In: *Proceedings of COLING 2012: Posters*. 2012, pp. 985–994.

[274] Juan C Reboredo and Andrea Ugolini. "Systemic risk in European sovereign debt markets: A CoVaR-copula approach". In: *Journal of International Money and Finance* 51 (2015), pp. 214–244.

[275] Rui Ren and Desheng Wu. "An innovative sentiment analysis to measure herd behavior". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* (2018).

[276] Carlos Rodriguez and Carlos Carrasco. *ECB Policy Responses between 2007 and 2014: a chronological analysis and a money quantity assessment of their effects*. Tech. rep. Financialisation, Economy, Society & Sustainable Development (FESSUD) Project, 2014.

[277] Ian Rogers. *The Google Pagerank algorithm and how it works*. 2002.

[278] Carlo Rosa and Giovanni Verga. "On the consistency and effectiveness of central bank communication: Evidence from the ECB". In: *European Journal of Political Economy* 23.1 (2007), pp. 146–175.

[279] Nouriel Roubini. "Why central banks should burst bubbles". In: *International Finance* 9.1 (2006), pp. 87–107.

[280] David Ruppert. *Statistics and data analysis for financial engineering*. Ed. by Springer. Vol. 13. Springer, 2011.

[281] Octavian Rusu et al. "Converting unstructured and semi-structured data into knowledge". In: *2013 11th RoEduNet International Conference*. IEEE. 2013, pp. 1–4.

[282] Krzysztof Rybinski. "A Machine Learning Framework for Automated Analysis of Central Bank Communication and Media Discourse: the Case of Narodowy Bank Polski". In: *Bank i Kredyt* 1 (2019), pp. 1–20.

[283] Shibley Sadique et al. "Soft information and economic activity: Evidence from the Beige Book". In: *Journal of Macroeconomics* 37 (2013), pp. 81–92.

[284] S Rasoul Safavian and David Landgrebe. "A survey of decision tree classifier methodology". In: *IEEE transactions on systems, man, and cybernetics* 21.3 (1991), pp. 660–674.

[285] Saygin Sahinoz and Evren Erdogan Cosar. "Economic policy uncertainty and economic activity in Turkey". In: *Applied Economics Letters* 25.21 (2018), pp. 1517–1520.

[286] Gerard Salton, Anita Wong and Chung-Shu Yang. "A vector space model for automatic indexing". In: *Communications of the ACM* 18.11 (1975), pp. 613–620.

[287] David S Scharfstein and Jeremy C Stein. "Herd behavior and investment". In: *The American Economic Review* (1990), pp. 465–479.

[288] Maik Schmeling and Christian Wagner. *Does Central Bank Tone Move Asset Prices?* Tech. rep. CEPR Discussion Papers, 2019.

[289] Helmut Schmid. "Part-of-speech tagging with neural networks". In: *Proceedings of the 15th conference on Computational linguistics*. Vol. Volume 1. Association for Computational Linguistics. 1994, pp. 172–176.

[290] Norbert Schwarz et al. "Ease of retrieval as information: another look at the availability heuristic". In: *Journal of Personality and Social psychology* 61.2 (1991), p. 195.

[291] Mark S Seasholes and Ning Zhu. "Individual investors and local bias". In: *The Journal of Finance* 65.5 (2010), pp. 1987–2010.

[292] Fabrizio Sebastiani. "Machine learning in automated text categorization". In: *ACM computing surveys (CSUR)* 34.1 (2002), pp. 1–47.

[293] John Sedunov. "What is the systemic risk exposure of financial institutions?" In: *Journal of Financial Stability* 24 (2016), pp. 71–87.

[294] Sunil Sharma and Sushil Bikhchandani. *Herd Behavior in Financial Markets; A Review*. Tech. rep. International Monetary Fund, 2000.

[295] Hersh Shefrin and Meir Statman. "Behavioral finance in the financial crisis: market efficiency, Minsky, and Keynes". In: *Rethinking Finance: New Perspectives on the Crisis*. Citeseer, 2011.

[296] Richard W Sias. "Institutional herding". In: *The Review of Financial Studies* 17.1 (2004), pp. 165–206.

[297] Julia Silge and David Robinson. *Text mining with R: A tidy approach*. Ed. by Oreilly&Associates Inc. "O'Reilly Media, Inc.", 2017.

[298] Eric Sims and Jing Cynthia Wu. "Evaluating central banks' tool kit: Past, present, and future". In: *Journal of Monetary Economics* (2020).

[299] Tara M Sinclair, Fred Joutz and Herman O Stekler. "Can the Fed predict the state of the economy?" In: *Economics Letters* 108.1 (2010), pp. 28–32.

[300] Amanpreet Singh, Narina Thakur and Aakanksha Sharma. "A review of supervised machine learning algorithms". In: *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. Ieee. 2016, pp. 1310–1315.

[301] Lee A Smales and Nicholas Apergis. "Does more complex language in FOMC decisions impact financial markets?" In: *Journal of International Financial Markets, Institutions and Money* 51 (2017), pp. 171–189.

[302] Neil R Smalheiser and Don R Swanson. "Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses". In: *Computer methods and Programs in Biomedicine* 57.3 (1998), pp. 149–153.

[303] M Nihat Solakoğlu, Ali Güvercin and Murat Engin Akkaş. "The Impacts of Elections and Central Banks Meetings on Herding Behavior: Evidence from Borsa Istanbul, 20". In: *Finans Sempozyumu* (2016), pp. 19–22.

[304] Marina I Solnyshkina et al. "Evaluating Text Complexity and Flesch-Kincaid Grade Level." In: *Journal of Social Studies Education Research* 8.3 (2017), pp. 238–248.

[305] Stuart Soroka, Lori Young and Meital Balmas. "Bad news or mad news? Sentiment scoring of negativity, fear, and anger in news content". In: *The ANNALS of the American Academy of Political and Social Science* 659.1 (2015), pp. 108–121.

[306] Tharsis Tuani Pinto Souza et al. "Twitter sentiment analysis applied to finance: A case study in the retail industry". In: *arXiv* (2015).

[307] Spyros Spyrou. "Herding in financial markets: a review of the literature". In: *Review of Behavioral Finance* (2013).

[308] Amit Srivastava and Francis Kubala. "Sentence boundary detection in Arabic speech". In: *Eighth European Conference on Speech Communication and Technology*. 2003.

[309] IMF Staff. "Herding in Financial Markets". In: *IMF Research Bulletin* 9.4 (2008).

[310] Stavros Stavroyiannis and Vassilios Babalos. "Time-varying herding behavior within the Eurozone stock markets during crisis periods". In: *Review of Behavioral Finance* (2019).

[311] L Venkata Subramaniam et al. "Business intelligence from voice of customer". In: *2009 IEEE 25th International Conference on Data Engineering*. IEEE. 2009, pp. 1391–1402.

[312] Aixin Sun and Ee-Peng Lim. "Hierarchical text classification and evaluation". In: *Proceedings 2001 IEEE International Conference on Data Mining*. IEEE. 2001, pp. 521–528.

[313] Don R Swanson et al. "Historical note: Information retrieval and the future of an illusion". In: *Journal of the American Society for Information Science* 39.2 (1988), pp. 92–98.

[314] Ann Taylor, Mitchell Marcus and Beatrice Santorini. "The Penn treebank: an overview". In: *Treebanks*. Springer, 2003, pp. 5–22.

[315] John B Taylor. "Discretion versus policy rules in practice". In: *Carnegie-Rochester conference series on public policy*. Vol. 39. Elsevier. 1993, pp. 195–214.

[316] Lillyn L Teh, Werner FM De Bondt et al. "Herding behavior and stock returns: An exploratory investigation". In: *Revue Suisse d'Economie Politique Et De Statistique* 133 (1997), pp. 293–324.

[317] Terry Therneau et al. *Package 'rpart'*. 2015. URL: `https://cran.r-project.org/web/packages/rpart/index.html`.

[318] Mark Thoma. "Bad advice, herding and bubbles". In: *Journal of Economic Methodology* 20.1 (2013), pp. 45–55.

[319] Ellen Tobback et al. "Belgian economic policy uncertainty index: Improvement through text mining". In: *International Journal of Forecasting* 34.2 (2018), pp. 355–365.

[320] Kristina Toutanova and Colin Cherry. "A global model for joint lemmatization and part-of-speech prediction". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Vol. Volume 1. Association for Computational Linguistics. 2009, pp. 486–494.

[321] Mohammed Musa Tumala and Babatunde S Omotosho. "A Text Mining Analysis of Central Bank Monetary Policy Communication in Nigeria". In: *CBN Journal of Applied Statistics* 10.2 (2019).

[322] Robert Tumarkin and Robert F Whitelaw. "News or noise? Internet postings and stock prices". In: *Financial Analysts Journal* 57.3 (2001), pp. 41–51.

[323] Andranik Tumasjan et al. "Predicting elections with Twitter: What 140 characters reveal about political sentiment". In: *Fourth international AAAI conference on weblogs and social media*. 2010.

[324] Nida Türegün. "Text Mining in Financial Information". In: *Current Analysis on Economics & Finance* 1 (2019), pp. 18–26.

[325] Alan Turing. "Computing machinery and intelligence-AM Turing". In: *Mind* 59.236 (1950), p. 433.

[326] Amos Tversky and Daniel Kahneman. "Loss aversion in riskless choice: A reference-dependent model". In: *The Quarterly Journal of Economics* 106.4 (1991), pp. 1039–1061.

[327] Geert Van Campenhout and Jan-Francies Verhestraeten. *Herding Behavior among Financial Analysts: a literature review*. Tech. rep. Hogeschool-Universiteit Brussel, Faculteit Economie en Management, 2010.

[328] S Vijayarani, Ms J Ilamathi and Ms Nithya. "Preprocessing techniques for text mining-an overview". In: *International Journal of Computer Science & Communication Networks* 5.1 (2015), pp. 7–16.

[329] Svitlana Voronkova and Martin T Bohl. "Institutional traders' behavior in an emerging stock market: Empirical evidence on polish pension fund investors". In: *Journal of Business Finance & Accounting* 32.7-8 (2005), pp. 1537–1560.

[330] Thierry Warin and William Sanger. "European Central Bank's monetary policy decisions: A dataset of two decades of press conferences". In: *Data in Brief* 20 (2018), pp. 794–798.

[331] Ivo Welch. "Herding among security analysts". In: *Journal of Financial Economics* 58.3 (2000), pp. 369–396.

[332] D White and RB Gramacy. *Package 'maptree'*. 2009. URL: `https://cran.r-project.org/web/packages/maptree/index.html`.

[333] Hadley Wickham et al. "Tidy data". In: *Journal of Statistical Software* 59.10 (2014), pp. 1–23.

[334] Hadley Wickham, Winston Chang and Maintainer Hadley Wickham. *Package 'ggplot2'*. 2016. URL: `https://www.tidyverse.org/`.

[335] Hadley Wickham and Garrett Grolemund. *R for data science: import, tidy, transform, visualize, and model data*. Ed. by "O'Reilly Media". "O'Reilly Media, Inc.", 2016.

[336] Hadley Wickham and Maintainer Hadley Wickham. *Package 'rvest'*. 2016. URL: `https://cran.r-project.org/web/packages/rvest/index.html`.

[337] Hadley Wickham and Maintainer Hadley Wickham. *Package 'stringr'*. 2019. URL: `https://cran.r-project.org/web/packages/stringr/index.html`.

[338] Hadley Wickham and Maintainer Hadley Wickham. *Package 'tidyverse'*. 2017. URL: `https://www.tidyverse.org/`.

[339] W John Wilbur and Karl Sirotkin. "The automatic identification of stop words". In: *Journal of Information Science* 18.1 (1992), pp. 45–55.

[340] Michael Woodford. *Central bank communication and policy effectiveness*. Tech. rep. National Bureau of Economic Research, 2005.

[341] Ho Chung Wu et al. "Interpreting tf-idf term weights as making relevance decisions". In: *ACM Transactions on Information Systems (TOIS)* 26.3 (2008), pp. 1–37.

[342] Sam Wylie. "Fund manager herding: A test of the accuracy of empirical results using UK data". In: *The Journal of Business* 78.1 (2005), pp. 381–403.

[343] Peter D Wysocki. *Cheap talk on the web: The determinants of postings on stock message boards*. Tech. rep. 98025. University of Michigan Business School Working Paper, 1998.

[344] Rui Xu and Donald Wunsch. "Survey of Clustering Algorithms". In: *IEEE Transactions on neural networks* 16.3 (2005), pp. 645–678.

[345] Gui-Rong Xue et al. "Deep classification in large-scale text hierarchies". In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 2008, pp. 619–626.

[346] Larisa Yarovaya, Roman Matkovskyy and Akanksha Jalan. "The Effects of a'Black Swan'Event (COVID-19) on Herding Behavior in Cryptocurrency Markets: Evidence from Cryptocurrency USD, EUR, JPY and KRW Markets". In: *Available at SSRN* (2020).

[347] Janet L Yellen et al. "Implications of behavioral economics for monetary policy". In: *Speech for the Federal Reserve Bank of Boston Conference:" Implications of Behavioral Economics for Economic Policy"*. Vol. 28. Sept. 2007.

[348] Marcia Lei Zeng. *Metadata*. Ed. by Inc. Neal Schuman Publishers. Neal-Schuman Publishers, Inc., 2008.

[349] Haiyi Zhang and Di Li. "Naïve Bayes text classifier". In: *2007 IEEE International Conference on Granular Computing (GRC 2007)*. IEEE. 2007, pp. 708–708.

[350] Wenbin Zhang and Steven Skiena. "Trading strategies to exploit blog and news sentiment". In: *Fourth international aAAI conference on weblogs and social media*. 2010.

[351] Yin Zhang, Rong Jin and Zhi-Hua Zhou. "Understanding bag-of-words model: a statistical framework". In: *International Journal of Machine Learning and Cybernetics* 1.1-4 (2010), pp. 43–52.

[352] Yongzheng Zhang, Rajyashree Mukherjee and Benny Soetarman. "Concept extraction and e-commerce applications". In: *Electronic Commerce Research and Applications* 12.4 (2013), pp. 289–296.

[353] Eric Zitzewitz. *Measuring Herding and Exaggeration by Equity Analysts and Other Opinion Sellers*. Tech. rep. Stanford University, Graduate School of Business, 2001.

[354] Feng Zou et al. "Automatic construction of Chinese stop word list". In: *Proceedings of the 5th WSEAS international conference on Applied computer science*. 2006, pp. 1010–1015.