
Universidad Carlos III de Madrid
Computer Science Department

Ph.D. Dissertation

# Financial Decision-Making Process Based on Unstructured Data Sources and Domain Ontologies

Author:

Mateusz Radzimski

Advisors:

Ángel García-Crespo
José Luis López Cuadrado

Leganés, 12 June 2017


UNIVERSIDAD CARLOS III DE MADRID

COMPUTER SCIENCE DEPARTMENT

PH.D. DISSERTATION

# FINANCIAL DECISION-MAKING PROCESS BASED ON UNSTRUCTURED DATA SOURCES AND DOMAIN ONTOLOGIES

MATEUSZ RADZIMSKI

ADVISORS:

ÁNGEL GARCÍA-CRESPO

JOSÉ LUIS LÓPEZ CUADRADO

Firma del Tribunal Calificador:

| | | Firma |
|---|---|---|
| Presidente: | Antonio Bibiloni Coll | _____ |
| Vocal: | Inmaculada Puebla Sánchez | _____ |
| Secretario: | Israel González Carrasco | _____ |

Calificación:

Leganés, 12 de Junio de 2017

*"Where is the Life we have lost in living?*

*Where is the wisdom we have lost in knowledge?*

*Where is the knowledge we have lost in information?"*

*– T.S. Eliot*

# *Acknowledgements*

First and foremost I want to thank all of you who made this journey more bearable and even enjoyable sometimes. I never meant to do a PhD, but life is unpredictable and takes you to unexpected places you have never imagined.

I dedicate this work to all the people who supported me during thse years. First of all: to my mom and dad, my girlfriend Magda and all my friends, whose support, advice and collaboration was so important to me. To Alejandro, Yuliana, Carlos, Jose Luis, Iván, Jose, Juanmi, Isra. To my advisors: Ángel and Jose. To all my Spanish friends. And to all of you who accompanied me in this journey. After all, it has been an unique experience and without you it would not be the same.

Thank you!

# *Abstract*

Nowadays a great number of financial decisions arrive from watching the information stream, selecting relevant data, analysing it and acting accordingly. With the increasing global competition, the need for swift data analysis, high accuracy and quality becomes a must.

For the majority of financial analysts, the main source for information is in the form of structured data. Such data can be easily processed and acted upon. However, there are vast amounts of knowledge that still can not be easily digested by computers, but have a great importance in our everyday life. For instance, (i) news are describing events and changes to the state of the world, (ii) columnists' opinions are providing arguments that are shaping our thoughts or (iii) experts' conclusions are influencing people decisions.

This thesis main objective is to employ unstructured data in the financial decision-making process, with the support of ontologies as the main backbone for knowledge representation. The whole financial-making process is contextualised in the scope of the Spanish market, where the main source of data is news and company disclosures published in the Spanish language.

The main contribution of this thesis is the creation of the *Decision Support System* (DSS) that follows a novel approach to incorporate unstructured data and domain (financial) ontologies into the automated financial decision-making process. Our approach employs Natural Language Processing (NLP) as means for extracting relevant information from unstructured sources. Moreover, semantics is applied thoroughly, not only in the process of information extraction but also in the knowledge modelling and the decision support.

# *Resumen*

Hoy en día un gran número de decisiones financieras son tomadas analizando el flujo de información, seleccionando los datos pertinentes, y finalmente actuando dependiendo del resultado de dicho análisis. Con el crecimiento de la competencia global, un análisis de datos rápido y con alta precisión y calidad se convierte en una necesidad. Para la mayoría de los analistas financieros, la fuente principal de información está formado por datos estructurados, tales datos se pueden ser procesados y gestionados fácilmente. Sin embargo, estos datos contienen grandes cantidades de conocimiento que todavía no pueden ser fácilmente digeridos por los ordenadores y tienen una gran importancia para los analistas, por ejemplo (i) las noticias, que describen eventos y cambios en el mundo, (ii) las opiniones de los columnistas, que proporcionan argumentos que están moldeando nuestros pensamientos o (iii) las conclusiones de los expertos, que influyen en las decisiones de las personas. El objetivo principal de la tesis es emplear datos no estructurados en el proceso de toma de decisiones financieras, con el apoyo de las ontologías como la principal columna vertebral para la representación del conocimiento. Todo el proceso de toma de decisiones financieras se contextualiza en el ámbito del mercado español, donde la principal fuente de datos son las noticias y divulgaciones publicadas en la lengua española. La principal contribución de esta tesis es la creación del Sistema de Soporte a la Decisión (DSS) que sigue un novedoso enfoque para incorporar datos no estructurados y ontologías de dominio (financiero) en un proceso automatizado de toma de decisiones financieras. El proceso emplea procesamiento de lenguaje natural (PLN) como medio para extraer información relevante de fuentes no estructuradas. Por otra parte,

la semántica se aplica a fondo, no sólo en el proceso de extracción de información, sino también en el modelado del conocimiento y el apoyo a la toma de decisión.

# Contents


1 *Introduction* 17

1.1 *Data-driven financial analysis* 18

1.2 *Unstructured data in the field of finance and economics — a big elephant in the room* 19

1.3 *Towards interoperable financial knowledge base* 23

1.4 *Objective of this thesis* 24

1.5 *Thesis organisation* 27

2 *State-of-the-Art* 29

2.1 *Information Extraction* 30

2.2 *Semantic modelling* 36

2.3 *Text analytics for decision support in finance* 51

2.4 *Conclusions* 58

3 *Research hypotheses & methodological evaluation approach* 61

3.1 *Hypotheses* 62

3.2 *Investigation process* 62

4 *Data acquisition for the analytical pipeline* 67

4.1 *Data source selection* 68

4.2 *Data acquisition* 69

4.3 *Data pre-processing* 72

4.3.1 *Meta data extraction* 72

4.3.2 *Boilerplate removal* 73

4.3.3 *Noise removal and normalisation steps* 76

4.4 *Corpus evaluation* 78

4.5 *Conclusions* 80

5 *Semantic modelling of high-level features* 83

5.1 *Foundations of the semantic model development* 84

5.2 *Relevant financial events* 87

5.3 *Common semantic event model* 92

5.4 *Taxonomies* 97

5.4.1 *Role taxonomy* 98

5.4.2 *Event type taxonomy* 100

5.5 *Event representation example* 100

5.6 *Note on the modelling approach* 103

5.7 *Conclusions* 105

6 *Financial events extraction* 107

6.1 *Feature analysis and corpus annotation* 108

6.2 *The bootstrapping technique* 111

6.3 *Information extraction pipeline for the knowledge base population* 113

6.4 Word representation in semantic vector space 115
6.5 Events classification with Convolutional Neural Networks 127
6.6 Comparing classical machine learning and neural network approach 129
6.6.1 Classical machine learning approach setup 130
6.6.2 "Deep" Neural Network approach 130
6.6.3 Hyperparameters and training details of CNN classifiers 131
6.6.4 Classification results 132
6.7 Conclusions 133

7 Decision-making based on unstructured data 135
7.1 Architecture for the Decision Support System 136
7.2 Model for decision support 139
7.3 Conclusions 141

8 Evaluation 143
8.1 Backtesting evaluation context 144
8.2 DSS model training 146
8.3 DSS model evaluation scenarios 149
8.4 Evaluation results 151
8.5 Validity of ALFREDO recommendations 155
8.6 Conclusions 156

9 Conclusions and future work 159
9.1 Conclusions 160

9.2 *Hypotheses validation summary* 162
9.3 *Thesis contributions* 163
9.4 *Future research* 165

*Bibliography* 167

*Appendices* 191

A *List of financial news sources* 193

B *Corpus statistics* 197

C *Ontology description* 201
C.1 *Event class taxonomy* 201
C.2 *Role taxonomy* 209

D *Detailed evaluation results* 229

# *List of Figures*


1.1 Rapid growth of the Digital Universe 20
1.2 Amount of data that is useful after being processed 21

2.1 The three main pillar of this thesis 29
2.2 Architecture of uimaFIR pipeline. 34
2.3 An example of interconnected triples forming a graph 38
2.4 NeOn methodology for building ontologies and ontology networks 42
2.5 A fragment of the Linked Open Data cloud diagram 43
2.6 The Linked Data Lifecycle, as defined in the project LOD2 46
2.7 Complete import graph of MUSING ontology. 48
2.8 Unstructured data used on a daily basis in financial decision-making process 54

3.1 Stages of the investigation process 63

4.1 Volume of acquired raw news documents 70
4.2 Typical webpage boilerplate elements 75
4.3 Identifying relevant content vs. the boilerplate text. 76
4.4 Boilerplate HTML code and extracted clean text. 77
4.5 Spotting corpus anomalies 78

5.1 Extending the event model 95
5.2 Two approaches to role modelling 98
5.3 Taxonomy of events 101
5.4 Semantic representation of events 103

6.1 Supervised training and classification 108

6.2 Manual sentence annotation for relation extraction task 109
6.3 Sentence with named entities annotated. 111
6.4 Overview of the extraction process. 115
6.5 Overview of the machine learning techniques. 117
6.6 Continuous bag-of-words model 120
6.7 Projection of words vectors into a space of two-dimensions 123
6.8 Projections of word vectors offsets for gender analogy 124
6.9 The architecture of the CNN classifier for event classification. 128
6.10 Training the CNN-Emb and CNN-Sem classifiers. 132

7.1 Overview of the news analytics for decision making 136
7.2 Architecture of the ALFREDO system 137
7.3 Hierarchcal model for decision support 140

8.1 Backtesting period 145
8.2 Event study parameters 147
8.3 Performance comparison of ALFREDO $T-1$ evaluation 154
8.4 QQ-plot for ALFREDO backtest results 157

B.1 Volume of acquired news documents per news site as a percentage of the whole corpus 198

# *List of Tables*


4.1 Overview of the news and company disclosures datasets 71
4.2 Corpus evaluation results 79

5.1 Classification of main financial events 88
5.2 Event model mappings 97
5.3 Ontology imports 97

6.1 Bootstrap seed examples 112
6.2 Word analogy examples 125
6.3 Results of different classification scenarios for single binary relation extraction. 133

8.1 Performance results for ALFREDO DSS 152
8.2 Summary of evaluation scenarios 153

B.1 Total number of acquired web documents per news site. 199

C.1 Event taxonomy class description. 201
C.2 Role taxonomy class description - top-level properties 210
C.3 Role taxonomy class description - unary relations 210
C.4 Role taxonomy class description - binary relations 224

# 1 *Introduction*

The aim if this chapter is to introduce the context of this thesis, and provide the reader with a convincing explanation of the motivation behind this work. Most importantly, we answer the question of why the matter at hand is relevant and how does it drive the purpose of this thesis.

The next three sections provide the context in a form of three emerging issues: (i) the necessity of including new data sources in the process of financial analysis, (ii) the growing body of available unstructured data and (iii) the difficulty in unleashing the potential of unstructured data due to the lack of interoperability and computational form.

As in the old saying: *necessity is the mother of invention*, we use the problem statement to drive the definition of the main objectives of this work. We introduce the principal contribution of this thesis and present how we plan to address the aforementioned issues.

## 1.1 *Data-driven financial analysis*

Nowadays a great number of financial decisions arrive from watching the information stream, selecting relevant data, analysing it and acting accordingly. With the increasing global competition, the need for swift data analysis, high accuracy and quality becomes a must.

For the majority of financial analysts, the main source for information is in the form of structured data. From reflective and accounting–based analyses to prospective, finance-based forecasts, from ratio-based fundamental analyses to time series technical analyses, they are all based on tables, numbers, and structured information, coming from various kinds of financial disclosures, market data, analyses, and more. Structured financial data gives many advantages, such as interoperability through domain-specific standards and common data formats, fosters the use of computational tools and automatic data processing.

Traditionally financial accounting data, be it Balance Sheet, Statement of Income, Cash Flow or other corporate disclosures, follow the basic assumptions of the Accounting Model. One of such classical and established accounting assumption is the *Transaction Approach*. It states that all reports in financial accounting record only events that (a) affect the financial position and (b) can be quantified in monetary terms (Gibson, 2012). However, many events that influence the future of the business entity are not necessarily reflected in any quantifiable form. As such events do not fall under the transaction approach they will not be captured by the structured data reports. For example, an announcement of an innovative product might increase the value of the company, while the information of an unexpected death of the CEO could make investors lose confidence, and in effect could worsen future prospects of the company.

This leads us to the first problem: not all the relevant information in accounting and finance is structured nor can be easily transformed into a structured form.

## 1.2 *Unstructured data in the field of finance and economics — a big elephant in the room*

The latest trends in financial reporting have the common principle objective of improving the efficiency of data exchange between business units, regulators and investors. For example, the XBRL standard for business and financial reporting proposes a structured, XML-based schema for publishing and interchanging financial reports (Bergeron, 2003). The XBRL as a way for publishing periodic financial statements of public companies is already adopted in the US and is being introduced in Europe as well (Pinsker and Li, 2008). The trend of giving the data structured form opens new horizons for interoperability between applications, and shifts the effort towards more high-level tasks such as financial analytics and decision-making process (Radzimski, Sanchez-Cervantes, et al., 2014).

Another problem with unstructured data is that it is not meant to have computable form because it is simply not practical to convert it into any agreed and standardised format. Financial news, analyses, public filings or other announcements are still being produced in natural language, textual form, by humans and to be consumed by humans. Providing the same data also in a computable form is not straightforward and usually economically unfeasible.

Unstructured data capture vast amounts of knowledge that still can not be easily digested by computers, but have a great importance in our everyday life. For instance (i) news are describing events and changes to the state of the world, (ii) columnists' opinions are providing arguments that are shaping our thoughts or (iii) experts' conclusions are influencing people decisions. Those are only a few examples of the significance of unstructured data in our everyday life. Moreover, the amount of data is growing at a fast pace (see Figure 1.1). According to market research company IDC, the total size of what is called a "Digital Universe" doubles every around 2 years, and by 2020 it will grow to about 44 trillion gigabytes (Gantz et al., 2012). Creating value from such data will increase,

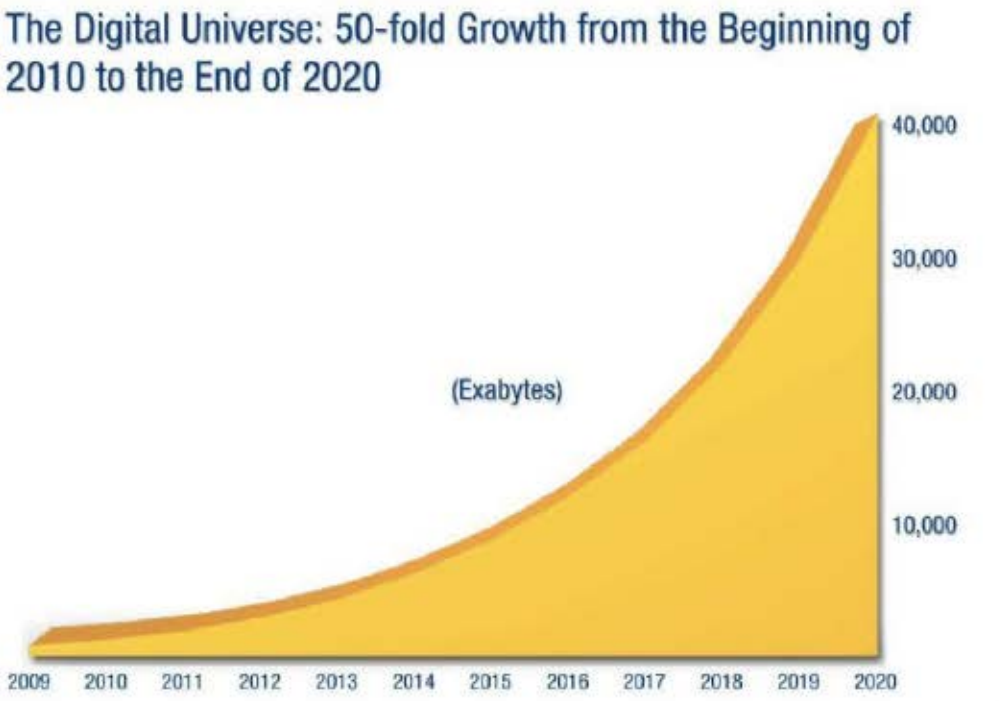



Figure 1.1: Rapid growth of the Digital Universe (Gantz et al., 2012).

and the amount of data that is useful if tagged and analysed will grow from 22% in 2013 to expected 37% in 2020, according to IDC forecast (see Figure 1.2) (Turner et al., 2014). The study underlines the problem of large quantities of data not having any form that could make it automatically processable. If we were able to attach (or extract) a relevant information to such data, it could be further analysed and employed, for instance, in a decision-making process. Otherwise, the value of such data is in a way sealed within its unstructured form — an unharvested opportunity waiting to be discovered.

As of now, the prognosis of the Digital Universe expansion holds true, and the increasing amount of data drives the growth of data analytics services. According to International Data Corporation (IDC), Big Data services and technology will grow at a 23% compound annual growth rate, reaching $48 billion in spending in 2019 (Nadkarni and Vesset, 2015). In the meantime, the cognitive capacity of the human brain remains on the same level. This means that humans will be less and less able to cope with such a huge amounts of ever-growing information. As


Figure 1.2: Amount of data that is useful after being processed (Turner et al., 2014).

a consequence, the information generated within our digital universe will have to be processed in an automatic and intelligent manner by machines, and those same machines will extract the relevant knowledge to be used by humans. The data monetization trend will drive the digital transformation initiatives and by 2020, 60% of information delivered to decision makers will be considered by them always actionable (Nadkarni and Vesset, 2015).

This development has already strong influence on *FinTech*[1] sector. The growing need to incorporate artificial intelligence in the industry is already a visible trend of the Digital Age, shifting the efforts from work-intensive to knowledge-intensive tasks. The process of making of financial decisions is frequently grounded in the knowledge that is coming from such data sources; therefore it will likely be disrupted by this trend as well.

One example of turning unstructured data into concrete value is analysis a great amount of opinions expressed in social networks about companies, their products, expectations about their results and future performance. Extraction of opinions (also called *sentiment analysis*) on a large scale, based on data from social networks such as Twitter, has already proved extremely useful in financial domain. Measuring sentiment in finance and economy is not a new concept at all. *Consumer confidence*

[1] FinTech stands for Financial Technology, an industrial sector consisting of companies aiming at disrupting financial services with technology.

indicators[2] that are based on a degree of optimism of consumers about the overall state of the economy are already widely used and proved valuable in macroeconomic analyses and decision-making process. They are important indicators to anticipate trends for the near future. The possibility of extending such approach to the world scale in order to automatically track and analyse sentiment about thousands of business entities, brands, and other aspects has been thoroughly studied with some promising results (Nassirtoussi et al., 2014).

This trend of automation of financial decision making is getting traction at this very moment, as we observe the breakout of the *Robo-Advisors*[3] market. It is estimated that automated portfolio management industry will account for 10% of all worldwide assets under management by 2020, which equates to around $8 trillion worth of assets. (Kocianski, 2016)

In this section, we have made a second important observation: unstructured data accounts for the majority of data produced by humans and contain vast amounts of relevant information for decision-making that are not easily available for automatic computation. The value contained in this data is still underexplored and there is a huge potential in unleashing it.

[2] A popular consumer confidence indicator is the "Consumer Confidence Index" (CCI), issued monthly by The Conference Board, an economic research organisation. It is based on a survey on 5000 households in the United States. In Spain, the CCI ("Indicador de Confianza de los Consumidores") is prepared and published by Centro de Investigaciones Sociológicas (CIS).

[3] Robo-Advisors are automated financial advisors that aim at managing assets portfolios relying on algorithms and artificial intelligence and minimising human assistance.

## 1.3 *Towards interoperable financial knowledge base*

While the structured form of the data gives many advantages such as easy programmatic access and syntactic data format, there are still various higher-level integration problems for financial data (Debreceny et al., 2010; O'Riain, Curry, et al., 2012). Furthermore, unstructured financial data is lacking *any* form that could make it interoperable or computable without extra human effort.

For instance, integrating data coming from heterogeneous sources, having only their syntactic representation implies additional work of aligning and "semantifying" used concepts. One of the widely successful methods is to align concepts with their semantic meaning, by "lifting" the data to a common semantic reference, where all concepts and relations have well defined and unambiguous meaning along with standard representation. Such approach is used in the Linked Data initiative. It is based on the standardised web semantic technologies and since its inception, it has been broadly adopted as a standard way for semantic data publishing, interlinking and querying.

The central part of semantic interoperability for financial knowledge base is the reuse of existing standards and domain knowledge, such as ontologies, vocabularies and taxonomies. This is especially important in the financial domain with numerous previous works in the fields of ontology engineering and semantic knowledge representation. The most important is the Financial Industry Business Ontology (FIBO)[4] — the business conceptual ontology standard that provides baseline structure and description for business entities, contractual and legal obligations, financial instruments, financial processes and market data (Bennett, 2013). FIBO is developed by the members of the EDM Council and standardised within the Object Management Group (OMG). While being backed by the industry, the ontology is relatively new, with several FIBO standards still under development at the time of writing of this thesis.

[4] FIBO overview can be found at `http://www.edmcouncil.org/financialbusiness`. Last accessed: 19/10/2016

That being said, we have learned that existing initiatives within the domain of knowledge representation provide standard and widely adopted means for sharing and interlinking financial knowledge. The rich preliminary work forms a well-grounded basis for the development of financial knowledge base model. The semantic technology offers a promising solution to bridge the gap between the unstructured data and the interoperability and computability required to unleash the full potential of financial data.

## 1.4 *Objective of this thesis*

In the previous subsections, we have made three major observations that constitute the main motivation for this work. Most of all, we learned that the unstructured data accounts of a vast majority of the fast-growing Digital Universe, and that the value of this data is still far from reaching its full potential. We showed that the data can have more value when it has more computable form, i.e. when is tagged or otherwise analysed, as shown in the Figure 1.2 (Turner et al., 2014).

From the entire Digital Universe, we are especially interested in unstructured data in the domain of Business Intelligence and financial decision-making in particular. We conclude that the value of such data can be unlocked when relevant information is represented in a way that can be further treated automatically. We look at semantic technologies as the enabler of data interoperability, through the formal knowledge representation and domain ontologies.

Therefore, **the main objective of this thesis** is to employ unstructured data in the financial decision-making process, with the support of ontologies as the main backbone for knowledge representation. The whole financial-making process is contextualised in the scope of the Spanish market, where the main source of data is news and company disclosures published in the Spanish language.

The main contribution of this thesis is the creation of the *Decision Support System* (DSS) that follows a novel approach to incorporate unstructured data and domain (financial) ontologies into an automated finan-

cial decisions-making process. The process employs Natural Language Processing (NLP) as means for extracting relevant information from unstructured sources. Moreover, semantics is applied thoroughly, not only in the process of information extraction, but also in the knowledge modelling and the decision support.

The whole process of Decision Support System creation can be divided into the three main parts:

- Analysis of the vast amount of unstructured data, identifying and extracting relevant features, and develop the complete analytical process for unstructured financial information extraction. The complete process should comprise all necessary steps, such as data acquisition, pre-processing, natural language processing, feature extraction and classification.
- Highly interoperable knowledge representation of the extracted data using semantic technologies and domain ontologies. The resulting knowledge base should support data publication, sharing, querying and analysis, in a way that supports data reuse, augmentation and integration with external datasets.
- Decision support model that incorporates extracted semantic knowledge base into analytical pipeline for supporting financial decision-making process.

The realisation of this process will be ALFREDO Decision Support System (ALFREDO stands for: AnaLysis of unstRuctured Data for Financial DecisiOn Making). The reason behind introducing ALFREDO is that it will encompass all the objectives into a coherent system. It will be also used for the final evaluation of the decision-making model.

From this perspective, we can now define **specific objectives** of this thesis:

**Objective 1:** Analysis of unstructured data sources from the point of view of feasibility for financial decision-making. Perform the acquisition of data and create a corpus for further analysis.

**Objective 2:** Identify most important information that can be obtained from unstructured data. Determine relevant features to be extracted.

**Objective 3:** Study semantic modelling in the domain of finance in order to provide a model for semantic knowledge base that represent financial information extracted from unstructured data.

**Objective 4:** Analyse extraction methods, and provide automatic classification of information for decision-making process.

**Objective 5:** Create model for decision support based on extracted features.

A significant part of the work of this thesis has been performed within the FLORA project, supported by Spanish Ministry of Industry, Tourism, and Commerce (TIN2011-27405). Some work has been inspired by author's participation in the European research projects: FIRST[5], MLi[6] and PHEME[7].

[5] Project FIRST overview: http://project-first.eu/. Last accessed: 24/04/2017

[6] Project MLi overview: http://mli-project.eu/. Last accessed: 24/04/2017

[7] Project PHEME overview: https://www.pheme.eu/. Last accessed: 24/04/2017

## 1.5 *Thesis organisation*

The remaining part of this thesis is structured as follows:

- Chapter 2 presents the state-of-the-art of the relevant fields of research that form the basis of this work.
- Chapter 3 describes the investigation process and **states the research hypothesis for this thesis**.
- Chapter 4 introduces the data acquisition process and corpora creation.
- Chapter 5 defines main information to extract from the text and later it focuses on semantic modelling for extracted data.
- Chapter 6 provides analysis of natural language processing methods to information extraction and describes the approach for financial facts classification.
- Chapter 7 explains the decision support model and describes architecture of the ALFREDO DSS.
- Chapter 8 evaluates the decision support model in the context of real world market data.
- Chapter 9 concludes the work and summarises the main outcomes of this work and outlines future research lines.

# 2 *State-of-the-Art*

The foundations of this thesis are grounded in the previous work in the following areas: (i) Information Extraction and Natural Language Processing (ii) Knowledge Representation and Semantic Modelling and (iii) Decision Support in the context of financial domain. Those three main pillars are illustrated in Figure 2.1. The following sections summarise the most relevant state-of-the-art for three aforementioned fields and perform an overview of existing systems for analysis of unstructured data in finance, their shortcomings and gaps; analysis of current investigation results on the effects of financial news on the stock market.

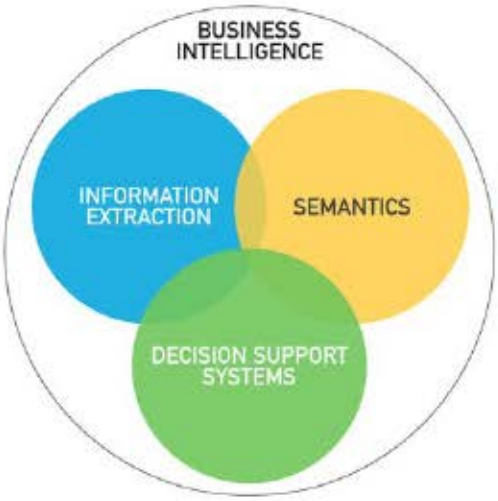



Figure 2.1: The three main research areas of this thesis with its context.

The main area of research of this thesis is in the scope of financial domain, more precisely in Business Intelligence (BI). In the inception of this term, BI was described as "the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards

a desired goal" (Luhn, 1958). Since then it evolved into a process of systematic acquisition, aggregation and analysis of information for decision-making. The most notable examples of such systems are Decision Support Systems (DSS), Executive Information Systems (EIS), Online Analytical Processing (OLAP), Data Warehouses and Data Mining systems (Chung et al., 2003). While the common application for BI systems is to analyse structured data (e.g.: sales, customer data, payment transactions, etc.) a great deal of investigations targets semi-structured and unstructured data. Moreover, unstructured and semi-structured data account for more than half of the overall amount of new data warehouse sources (Russom, 2007). Incorporating unstructured data into BI systems heavily involves such fields as Information Extraction (IE) and Natural Language Processing (NLP) in order to understand data and transform it into machine-readable formats for further processing.

## 2.1 *Information Extraction*

Transforming human-readable texts into machine-readable documents attracted significant attention since the mid-20th century. Natural Language Processing techniques have been invented to automatically convert unstructured data into formats that allow further analysis. The possibility of automatically analyse huge amounts of texts in a search for relevant data became a popular use case in many domains, such as finance, insurance, medicine, transportation, only to name a few.

Information Extraction (IE) is one of the most important fields of NLP. Its goal is to extract relevant information from unstructured and semi-structured data. Those can be textual sources, but also audio, video or photos. However for the purpose of this thesis, we will further focus on textual sources exclusively. Typically, tasks involving extraction of some facts are performed by analysts or highly specialised personnel. For example a physicist reading patient's health record, or financial analyst digging through company reports in order to assess the risk of investment. In many cases, it is a time-consuming process, especially when very big number of documents is involved. This is where IE systems

excel. They can automate such tasks to a great extent and very quickly process input textual documents in order to find and extract relevant data. IE systems can also serve as a first step for developing more complex artificial intelligence systems.

In this sense, Information Extraction goes way beyond a simple Information Retrieval (or Document Retrieval), which is typically limited to keyword-based document classification (Goodrum, 2000). A prominent class of Information Retrieval systems are web search engines. On the other hand, Information Extraction systems employ more advanced techniques by performing textual analysis and very often they employ domain knowledge bases (such as ontologies) and artificial intelligence in order to support more sophisticated use cases.

From the high-level point of view, IE system can be divided into two groups according to the knowledge engineering method (Sebastiani, 2002):

- Rule-based Information Extraction systems use knowledge expressed in a form of domain-specific rules. Domain experts engineer rules manually in order to match and extract information, based on specific cases and patterns that appear in the input text. The rules are typically logical expressions in the form of if-then assertions or conditions. It takes significantly more time to create such systems and it is a very laborious process, but they can reach very high precision and accuracy once developed. Rule-based systems are also a preferred way for constructing IE systems when domain taxonomies, lexicons or ontologies are already available.
- Machine learning systems are based on trained algorithms to classify and extract information. In the supervised learning, a corpus of human-annotated documents is used to train the extraction model. Given enough cases, the algorithm is further able to automatically classify new data on its own. While the preparation of the corpus needs also human assistance, exist techniques to improve and shorten the whole process (e.g. through distant supervised learning). Also, it is generally less laborious to develop a machine-learning system.

However, precision of such system depends on many factors, such as concrete scenario and domain, algorithms used, size and quality of the training set. For specific use cases exist semi-supervised and unsupervised learning methods (e.g. clustering). Machine learning IE systems require training data and is usually preferred when such data is easily available.

On the intersection of those two methods, there is a plethora of hybrid approaches and ensembles that combine both rule-based and machine learning approaches in order to leverage advantages of both groups. Such systems typically yield very good results, for instance using machine learning for fast classification, and later refining the results with rule-based systems in order to improve overall accuracy.

The Information Extraction domain gained its impetus in the late eighties by the series of Message Understanding Conferences (MUC) organised by Defense Advanced Research Projects Agency (DARPA). It was structured in topics (starting with military topics, and slowly shifting towards more general themes), and the goal was to extract and fill necessary information on a certain topic based on the input texts, an activity called slot filling. The MUC conferences contributed to the development of IE and set common standards for comparison and evaluation of IE algorithms by the means of precision and recall and established common tasks within the IE, such as named entity recognition, co-reference resolution or relationship extraction. Nowadays the most important NLP and IE activities are performed within the Conference on Natural Language Learning (CoNLL) that sets special tasks every year. Along with the task (called "shared task"), the training and testing corpus is published so the participants can evaluate their systems and compare results. Over last years, shared tasks significantly advanced many areas of NLP and improved the state-of-the-art in the Information Extraction field (Grishman and Sundheim, 1996).

Linguistic analysis of texts consists of specific tasks that are usually performed, starting from the most basic (that operate on character and symbols level) to more complex (operating on words, sentences and se-

mantics). For instance a sentence splitting task must be accomplished in order to properly parse a sentence, while more complex models typically require a presence of multiple text features (Jurafsky and Martin, 2000). The most common steps in many text analysis tasks are the following:

- Tokenisation is a process of transforming strings of characters into chunks called tokens. They typically represent single words, but can also represent numbers, compound words, etc. Tokenisation also takes into account punctuation.
- Sentence splitting is a process where individual sentences are separated from the stream of tokens.
- Part-of-speech tagging (POS tagging) is a syntactic analysis of text and identifying the part-of-speech of words, e.g. verbs, nouns adjectives, etc.
- Dependency parsing is an analysis of the grammatical relation of words in a sentence and creating a structure that represents it in a form of a dependency (or constituency) tree.
- Named Entity Recognition (NER) is detecting named entities (e.g. people, places, organisations, etc.) and tagging them in a text.
- Co-reference resolution is the identification of all words that refer to the same object or person (e.g. linking pronouns to actual person they refer to). This can be done also in a cross-sentence manner.

The given list shows most typical tasks for text annotations but is not exhaustive, as more high-level features can be extracted from texts.

Components realising those tasks are integrated into bigger frameworks that allow for more complex IE jobs. General Annotation for Text Engineering (GATE) (Bontcheva et al., 2004; Cunningham, 2002) is a very popular framework for constructing custom NLP pipelines for text processing. The core of GATE is A Nearly-New Information Extraction (ANNIE) system that offers various components for text processing. GATE allows for plugging ontologies and gazetteers in order to perform

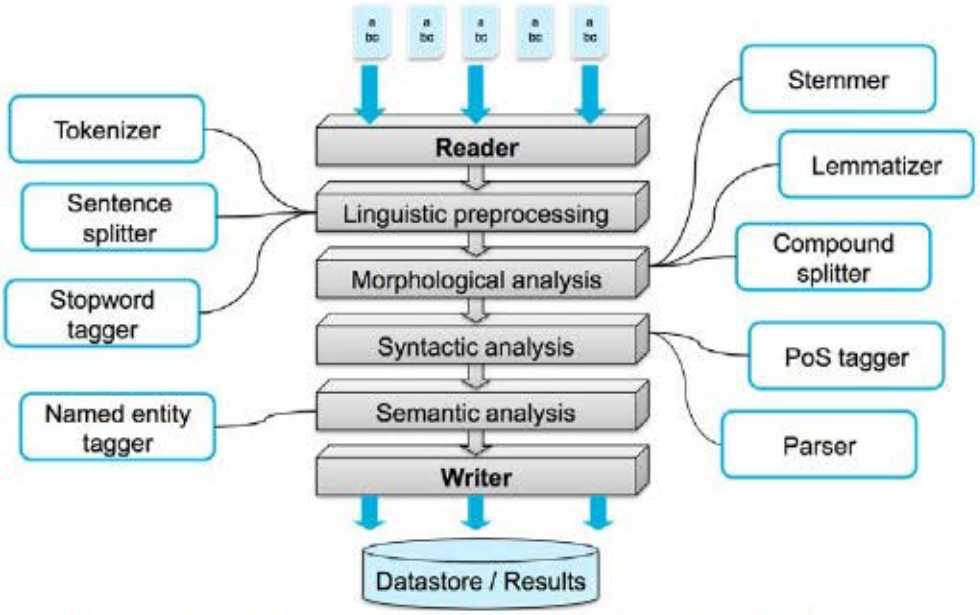



Figure 2.2: Architecture of uimaFIT pipeline (Eckart de Castilho, 2013).

ontology-based Named Entity Recognition and concept matching. The Java Annotation Patterns Engine (JAPE) rule-based language allows for user-defined transducers on top of extracted features, which can be useful for e.g. defining grammar rules or custom information extraction patterns.

Another very popular framework is the Apache UIMA (Ferrucci and Lally, 2004). It is component-based architecture for analysis of unstructured data. in its core lies a flexible text annotation type system called Common Analysis Structure (CAS) and common API for analysis components. Text processing with UIMA works like an assembly line, where each component is adding new annotations and refinements to the previous steps. uimaFIT system is built on to of that and provides facilities for creating analytical pipelines. UIMA Ruta is another addition that brings the pattern matching language over UIMA annotations. The UIMA framework does not provide its own models for performing NLP tasks but allows reusing other well-established ones. E.g. the DKPro project (Eckart de Castilho and Gurevych, 2014a) aims at integrating useful models and components coming from other projects and frameworks, but preserving the flexibility of uimaFIT pipelines. Figure 2.2 gives an

overview of the uimaFIT pipeline architecture. It can be extended to contain extra functional blocks for further analyses.

There are numerous other NLP frameworks available, with their own features and advantages. The most notable are: CoreNLP (Manning et al., 2014), OpenNLP[1], NLTK (Bird et al., 2009), Mate-tools (Björkelund et al., 2010), NLP4J (previously known as ClearNLP) [2]. They mostly come with their own components and models for natural language processing. Most of the models they contain can be used from within DKPro framework in uimaFIT pipelines.

In the recent years, the state-of-the-art in NLP techniques has been substantially improved thanks to the use of "deep" neural networks. This became a significant breakthrough for machine learning and artificial intelligence in general. Since then it widened the potential of natural language processing and started to bring it even closer to the real-world scenarios (Goth, 2016).

Traditional machine learning approaches to text analysis usually treat words as tokens, without any notion of its semantics, so each word is, in the end, becoming a different number regardless their potential similitude. Some operations such as word stemming or lemmatisation improve this situation by reducing all inflected forms into its corresponding lexeme. While this is usually enough to account for grammatical variations of the word, it is not sufficient to relate words that are similar in meaning.

Two modern approaches were proposed to bridge this gap and improve this situation. In (Mikolov, Corrado, et al., 2013) author proposes word2vec approach for learning word embeddings that capture semantic relationship between words, by representing words as vectors in high-dimensional vector space. Word2vec word embeddings are created by predicting word based on its context. A similar goal was achieved in (Pennington et al., 2014) by analysing word co-occurrences frequencies.

[1] The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text. URL: https://opennlp.apache.org, Last accessed: 18/04/2017

[2] NLP tools developed by Emory University. URL: https://github.com/emorynlp/nlp4j, Last accessed: 18/04/2017

Authors discovered that probabilities of words co-occurrences encode some meaning, where related words co-occur more often than those unrelated ones. As a consequence, semantically similar words have vectors that are also closer to each other in the high-dimensional vector space.

Another important improvement was the breakout of deep learning (LeCun, Bengio, et al., 2015) and new methods for faster training of multilayer neural networks, more suitable architectures for language processing and new approaches to language modelling with neural networks (Collobert et al., 2011). This enabled the new class of artificial intelligence systems that pushed state-of-the-art of natural language processing even further. The most prominent examples are the use of a family of Convolutional Neural Networks (CNN) for sentence modelling (Hu et al., 2014; Kalchbrenner et al., 2014) and classification (Kim, 2014), relations classification and extraction (Nguyen and Grishman, 2015; Zeng et al., 2014a), language modelling on character-level with Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) (Kim et al., 2016; Sundermeyer et al., 2015). What they have all in common is that it is possible to build an end-to-end natural language classifiers by only using neural networks architectures. These new advancements only scratched the surface of new possibilities for information extraction and natural language processing to transform textual sources into actionable knowledge.

## 2.2 *Semantic modelling*

With the dynamic development of the World Wide Web and exploding amount of data suddenly available to anyone, it became clear that the information is not sufficient when it is separated from its meaning (Berners-Lee et al., 2001). The possibility of representing the knowledge so that computers can understand and act upon it could bring many benefits in the future that even now are hard to predict. Enabling automatic processing of information with well-defined meaning could lower the barriers for performing more advanced tasks, such as decision-making process.

Sharing knowledge between different participants requires establish-

ing a common understanding of the given domain and agreeing on the meaning of the concepts involved. This implies creating formal definitions for entities, their properties, relationships and axioms that govern a specific domain (Hendler, 2001). Such definitions that encapsulate the knowledge about a particular domain are called ontologies. They play a central role in knowledge sharing, for they represent an accepted formal specification of a shared conceptualisation (Gruber, 1995b), where shared "reflects the notion that an ontology captures consensual knowledge, that is, it is not private of some individual, but accepted by a group" (Studer et al., 1998).

In the past decade, ontology engineering and constructing a semantic representation of information was an important topic of research in artificial intelligence. The flagship idea of extending the World Wide Web with metadata in a form of semantically linked concepts, relationships and properties, has driven the development of semantic technologies under the common umbrella term of *Semantic Web* (Berners-Lee et al., 2001). In this sense, the new Web would not only contain syntactic information, designed and interpreted for humans (e.g. HTML), but also semantic information, that could be interpreted by computers (e.g. RDF[3]). The backbone of the Semantic Web would be shared ontologies providing a common reference for concepts and their meaning (Davies et al., 2003).

Semantically describing a resource means to provide metadata[4] — an additional information about a resource. Such metadata gives machines a hint about the object at hand, its properties, characteristics and relationships with other resources, while resources can be any arbitrary thing: a person, a document, physical or abstract objects and more (Schreiber and Raimond, 2014). This allows machines to "understand" what the object is, and how it can be automatically processed.

Resource Description Framework (RDF) is a framework for representing information on the Web (Cyganiak et al., 2014). The core idea is that all statements are represented in a form of *triples* consisting of three

[3] Resource Description Framework, more on that later

[4] Metadata or meta information is information about information

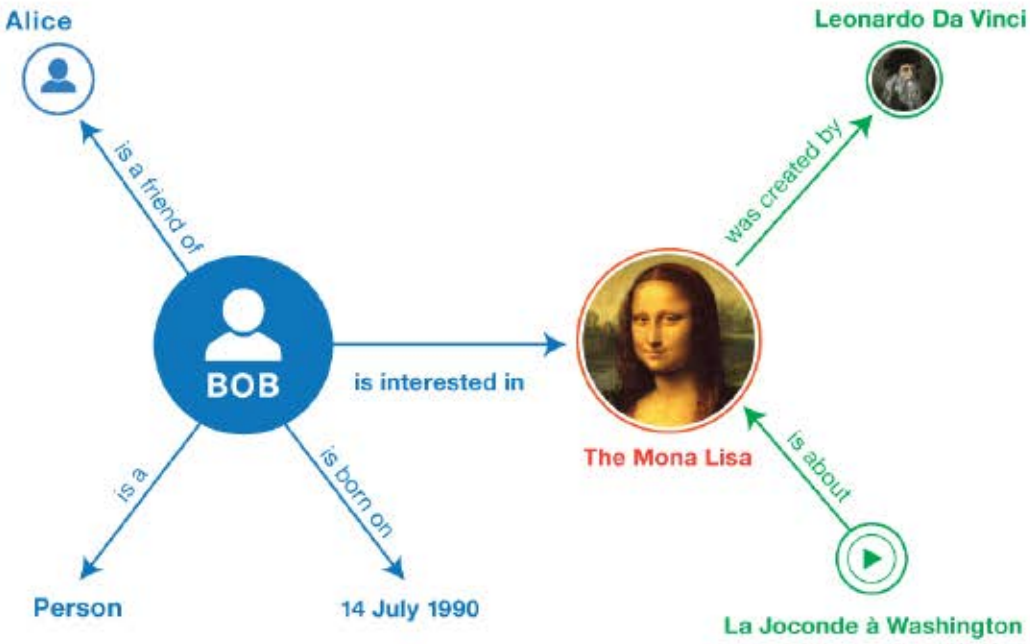



Figure 2.3: An example of interconnected triples forming a graph (Schreiber and Raimond, 2014).

parts: a *subject*, a *predicate* and an *object*. Subject is a resource of interest that is being described and is represented with an URI[5]. Predicate is a property defining a concrete aspect being described: an atribute, characteristics or relation with other resource. Object is a value of the defined property. It can be either a concrete value (a literal, number, date, etc.) or other resource (described with an URI). Due to the fact that object of one triple can be subject of another triple, such triples can be connected with each other and represent a graph[6]. Figure 2.3 presents a graphical example of triples related with each other and forming a graph. Resources are represented here with circles, predicates with edges and some objects are literal values (e.g. "Person" or "14 July 1990"). For instance in the following triple (*Bob, is interested in, The Mona Lisa*), *Bob* is a subject, *is interested in* — a predicate and *The Mona Lisa* — an object. But then, *The Mona Lisa* is a subject for another triple (*The Mona Lisa, was created by,*

[5] URI stands for Universal Resource Identifier

[6] RDF datasets are therefore often called RDF graphs.

*Leonardo Da Vinci*). Providing standard URIs for resources allows for an integration of different datasets (RDF graphs) by merging them based on the same URIs. The RDF data model can be serialised into different human or machine-readable formats, such as XML (the canonical RDF/XML format) but also into Turtle formats (N-Triples, Turtle, TriG and N-Quads), RDFa and JSON (JSON-LD[7]).

RDF in very useful language for describing things on the web, but lacks expressiveness when it comes to creating taxonomies and schemas for metadata. RDF Schema (RDFS) was proposed to fill this gap and provide a data modelling vocabulary for RDF (Brickley and R. Guha, 2008). Semantically RDFS is an extension to RDF, adding new constructs for defining vocabularies, relations of properties, classes, more detailed characteristics of resources, collections and reification methods among others.

A next step forward towards more expressive and full-fledged language for constructing ontologies is the Web Ontology Language (OWL). OWL was born on previous efforts in ontology modelling, such as Ontology Interchange Language (OIL) (Fensel et al., 2000) and DAML+OIL (DARPA, Agent Markup Language + OIL) (Horrocks, 2002). OWL consists of a family of languages (profiles) with a formal definition of meaning and used for defining ontologies. The OWL languages are canonically represented in RDF, extending syntax in order to include more expressive constructs, such as new classes and distinction between classes and individuals, more complex property characteristics, restrictions, classes intersections and more (W3C, 2012). OWL also allows for more powerful reasoning, not only rule-based but also using elements of description logics. The latest version is OWL 2 and it is built on top of previous revisions, adding three new profiles for dealing with the trade-off between expressiveness and computational complexity (Motik et al., 2012):

- OWL 2 EL: designated specifically for ontologies with large number of properties and classes, provides expressive power for this class of

[7] More details on representing RDF in JSON-LD format can be found in (Sporny et al., 2013) in Chapter 9: Relationship to RDF.

ontologies and guarantee decidability in polynomial time with regard to consistency, class expression subsumption and instance checking. This profile however restricts e.g. negation and disjunction.

- OWL 2 QL: particularly aimed at ontologies with large amounts of instance data and oriented towards query answering tasks. For this reason, OWL 2 QL can be implemented with relational database systems. This profile is more limited in terms of expressivity, in return for good performance for query answering.

- OWL 2 RL: targeted at those applications that need efficiency and scalable reasoning, despite the cost of lower expressivity. This profile is restricted only to those assertions that can be implemented with rule-based reasoner.

SPARQL Protocol and RDF Query Language (SPARQL) is another member of the family of Semantic Web languages (Prud'hommeaux and Seaborne, 2008). SPARQL is used to query semantic knowledge bases, in a similar way that SQL is used for relational databases. Queries are performed by graph patterns matching over RDF repositories by using the following constructs: simple triple patterns, conjunctions, disjunctions, and optional patterns. The result of the query can be either: a (i) set of values bound to variables in a query pattern match (ii) RDF graph created by substituting variables in graph template (iii) boolean value by testing if a pattern matches the graph (iv) RDF graph describing a particular resource. The semantics of SPARQL is defined by RDF and recommends Turtle format for expressing graph patterns.

Development of ontologies for different aspects of information sharing was a key point in bringing the Semantic Web vision forward. In its inception being more an art rather than an engineering discipline (Gómez-Pérez, Fernández-López, et al., 2004), ontology development required more structured approach. A methodology was needed for ontology modelling, which could provide guidelines and support that cover the whole life cycle: from creation and refinement to maintenance and sharing.

The earliest approaches to defining ontology engineering approach were performed by Lenat and R. V. Guha (1990) on creating the Cyc ontology and by Fox (1992) in the project TOVE, by applying knowledge engineering in the enterprise modelling domain. Since then, further methodologies were proposed, of which the most complete are METHONTOLOGY (Blazquez et al., 1998; Fernández-López et al., 1997), On-To-Knowledge (Staab et al., 2001), DILIGENT (Pinto et al., 2004) and NeOn methodology (Suárez-Figueroa et al., 2012). The complete comparison of most important methodologies can be found in (Fernández-López, 1999). The NeOn methodology is noteworthy as it grounded in experience gained by its predecessors. It sets nine most plausible scenarios for developing ontologies, well suited to different project requirements and life cycles, as there is no single approach that satisfies all development needs (Brooks, 1995; Pfleeger, 2009). NeON methodology emphasise the reuse of ontological and non-ontological resources, reengineering and merging of ontologies, takes into account collaboration and dynamism and finally sets the common glossary of process and activities for ontology development cycles (Gómez-Pérez and Suárez-Figueroa, 2009).

Figure 2.4 shows an overview of the NeOn methodology activities and associated scenarios: (1) from specification to implementation, (2) reusing and reengineering non-ontological resources, (3) reusing ontological resources, (4) reusing and reengineering ontological resources, (5) reusing and merging ontological resources, (6) reusing, merging and reengineering ontological resources, (7) reusing Ontology Design Patterns (ODPs), (8) restructuring ontological resources, (9) localizing ontological resources.

Ontology development as a preliminary step in building worldwide Semantic Web proved to be a rather difficult task and the community became disillusioned with slow adoption. In its primary form, Semantic Web did not live up to everyone's expectations and did not achieve its grandiose objectives. Ontologies themselves had not become ubiquitous on the web and semantic resources were scarce. Moreover, performing

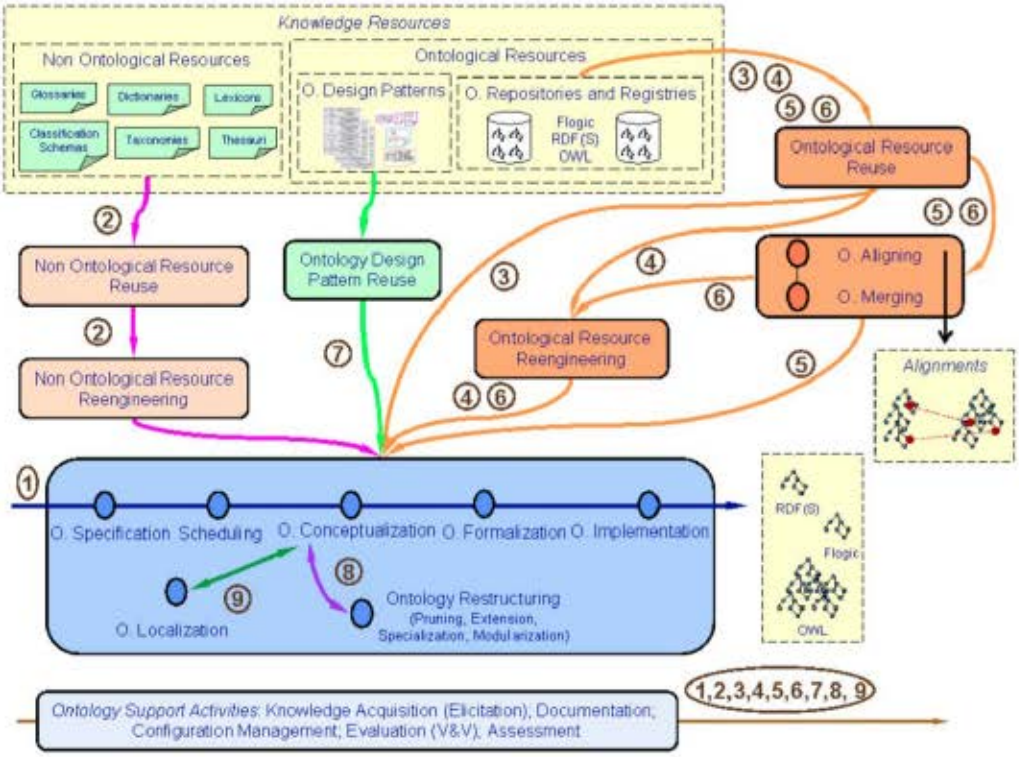



Figure 2.4: Resources (dashed boxes), activities and processes (solid boxes) and scenarios (arrows and numbers) of NeOn methodology for building ontologies and ontology networks (Gómez-Pérez and Suárez-Figueroa, 2009).

large-scale reasoning was too costly relative for benefits it could provide.

Although the idealistic vision of the early visionaries of the "Semantic Web" has never been fully achieved, it helped to create strong foundations for knowledge reuse standards and technologies. One of the most successful initiatives of the Semantic Web, the Linked Open Data (LOD) cloud, and its most renowned project DBpedia[8] became a reference for representing and sharing information on the Web (Bizer et al., 2009). DBpedia uses Semantic Web technologies and Linked Data principles to provide a semantic knowledge base of over 1.46 billion facts and 10 million additional concepts that are available for navigating and querying (Lehmann et al., 2015). It also became the most important hub for data interlinking and integration of different datasets through the reuse of

[8] The DBpedia project is available at `http://dbpedia.org/about`. Last accessed: 24/04/2017

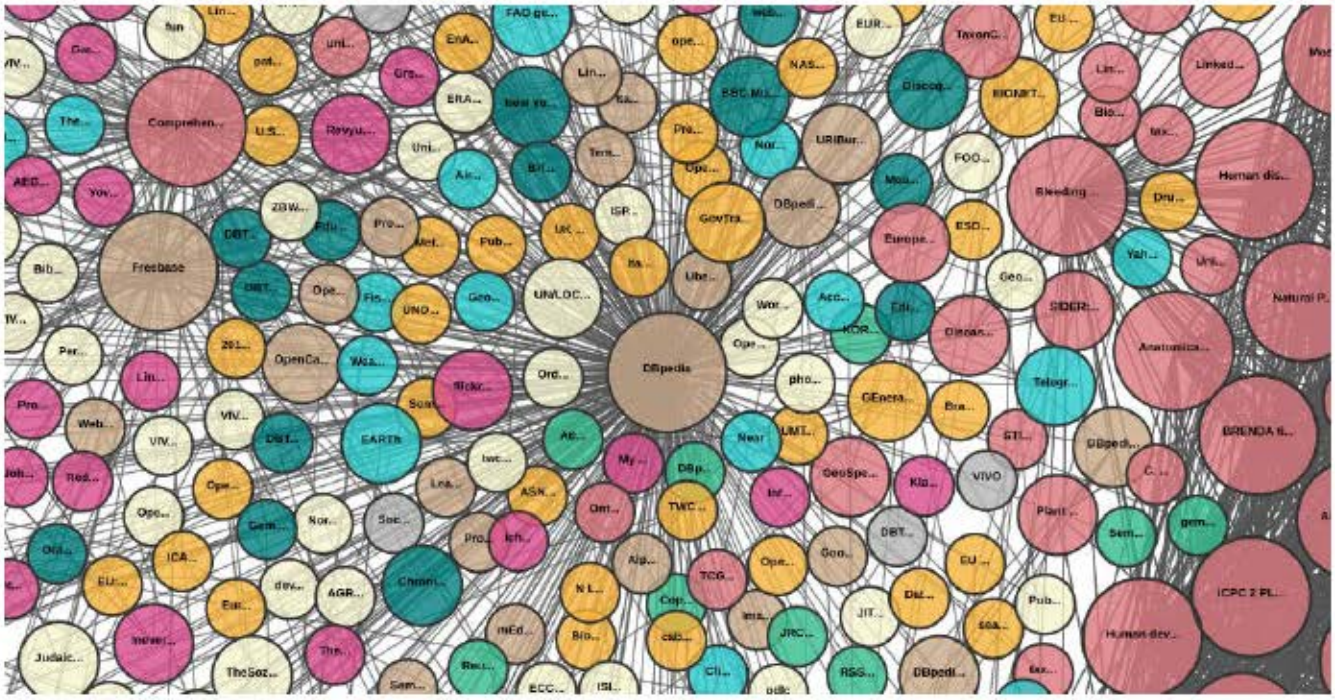



Figure 2.5: A fragment of the Linked Open Data cloud diagram as of February 2017. DBpedia is the central dataset of the LOD cloud. The complete diagram is maintaned by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak at `http://lod-cloud.net/`. Last accessed: 18/04/2017.

common concepts defined already in DBpedia. Figure 2.5 shows a fragment of LOD cloud, a diagram presenting Linked Data datasets published and available online, and their interconnections. Concepts that are common across different datasets are linked together (through the `owl:sameAs` relation). This allows for cross-dataset navigation, querying and reasoning. The central dataset of the LOD cloud is the DBpedia, as the main source of general top-level concepts, properties and hierarchies. Other semantic datasets are linking to DBpedia in order to establish a common understanding of their concepts.

In the core of the Linked Data approach for data publication there are four rules, as defined by Berners-Lee (2010):

1. Use URIs to name and identify things.
2. Use HTTP URI so that these things can be looked up.
3. Provide useful information about what a name identifies when it's

looked up, using open standards such as RDF, SPARQL[9], etc.

4. Refer to other things using their HTTP URI-based names when publishing data on the Web.

The Linked Data approach evolved into comprehensive and standard way for publishing public and open data. The process of data publication has been defined in multiple ways, e.g. by Hyland, Hausenblas and Villazón-Terrazas in (Hausenblas et al., 2011), with the most up-to-date methodology performed within the EU project LOD2. The LOD2 life cycle has been presented in the Figure 2.6. The different stages are the following (Auer, Bühmann, et al., 2012):

- Extraction: On many occasions data is represented in a way that does not facilitate an easy access, computation or querying. Extracting such data is a first step in transforming it into a Linked Data.
- Storage and Querying: Semantic data storage is very different from traditional databases. Handling graph representation, inference, querying, distributed cloud storage are some of the issues addressed in this stage.
- Authoring: creating new rich semantic knowledgebases in an intuitive and not overcomplicated ways, embracing social collaborative tools and techniques, such as Semantic Wiki that lowers the threshold for human participation in creating semantic datasets.
- Interlinking: making relations between semantic datasets is crucial for the growth of the Linked Open Data cloud. Tools for human-assisted semi-automatic linking of semantic concepts foster the cross-dataset knowledge integration.
- Classification: linking existing raw concepts to their corresponding classes in semantic vocabularies and ontologies allows "lifting" them and give them semantic meaning.

[9] SPARQL Protocol And RDF Query Language, https://www.w3.org/TR/rdf-sparql-protocol/. Last accessed: 24/04/2017.

- Quality Analysis: assessing LOD datasets from the point of view of the quality of data can be expressed with regard to different characteristics, such as coverage, structure and provenance.

- Evolution: LOD datasets evolve to adapt to the ever changing web. Facilitating evolution and repair of datasets, while keeping stable URIs, traceability of changes and semi-automatic guidance in knowledge base restructuring is the goal for tools supporting this stage.

- Search, Browsing and Exploration: for end-user it does not matter what exactly is under the hood as long as she can successfully navigate the data. Taking advantage of the nature of data, exploration can be facilitated with different techniques such as faceted browsing, geographical visualisations, aggregated views, etc.

The circular shape of the Linked Data Lifecycle diagram denotes that those stages are not isolated from each other but rather influence themselves and are interdependent. A change in one stage can trigger necessary actions in another.

Every stage is supported by its corresponding tools that aid users to deal with concrete problems at each point of the life cycle. The necessary tooling has been also covered by the LOD2 project in a form of the LOD2 Stack (Van Nuffelen et al., 2014).

The Linked Data approach, as presented above, is successfully used to publish data in an interoperable way that facilitates dataset interlinking and semantic representation of financial concepts. It is also employed in the field of finance and economy, enabling new scenarios of data integration (O'Riain, Harth, et al., 2012), querying and analysis (Radzimski and Sánchez-Cervantes, 2012).

Semantic financial modelling plays important role in transforming unstructured or otherwise unactionable financial data into computable, shareable and navigable datasets. Several ontologies were constructed to provide semantic foundation for expressing financial data.

In (Vanderlinden, 2011) author develops "Finance Ontology" that contains financial instruments, parties involved in financial transactions,

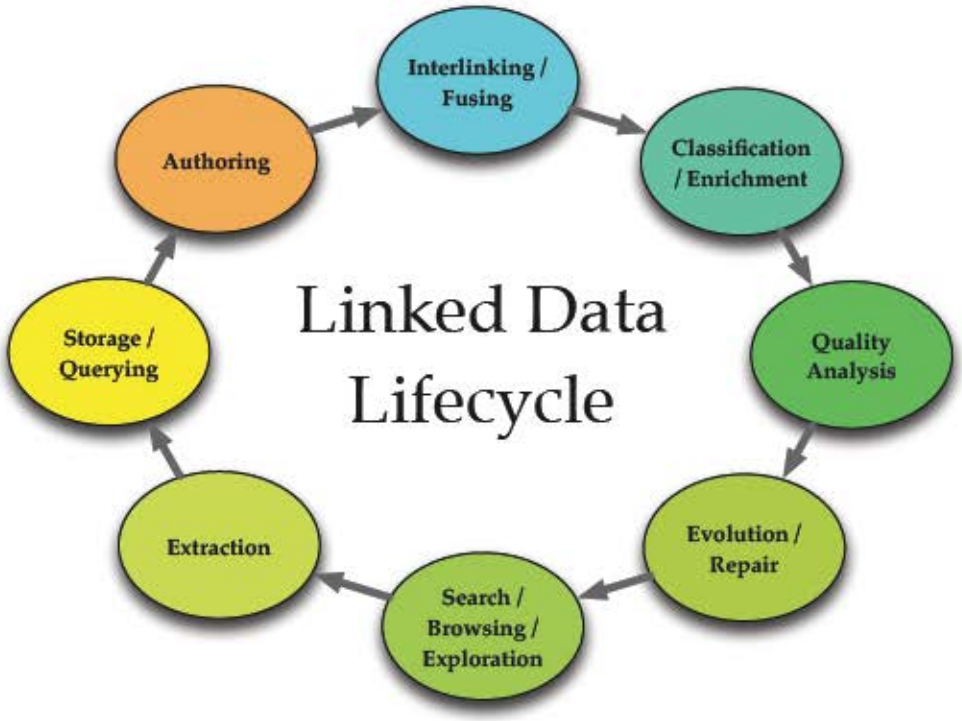



Figure 2.6: The Linked Data Lifecycle, as defined in the EU FP7 project LOD2.

markets, currencies, economic activities. It also borrows from ISO 10962 financial instruments classifications and ISO 20022 universal financial industry message scheme. Therefore it is mostly oriented at modelling securities and related activities.

Another ontology was proposed by Lara et al. (2006) and targets investment funds. The main idea is to use common vocabularies and concepts to gather and integrate historical information on investment fund from disparate and heterogeneous data sources. Authors begin with defining a taxonomy based on established XBRL format and define a mapping from XBRL into OWL ontologies. In consequence, the availability of commonly shared models can improve the process of data integration. This can bring an added-value and foster investment funds analyses.

Another branch of knowledge modelling in finance targets fraud prevention and detection. The European project FF POIROT aimed at developing an ontology-based system for fraud and scam detection in elec-

tronic communications (Kerremans et al., 2005; Zhao et al., 2004). A similar approach for detecting illegal solicitation of financial products was followed in (Gao and Zhao, 2005) by developing an application-specific ontology for linguistic processing of emails. On the other hand, Zaki and Theodoulidis (2013) describe ontology development for financial market monitoring and surveillance. The ontology captures various fraud practices and is used as the backbone for information extraction from textual sources. It is validated in three scenarios: (i) extracting financial fraud concepts from Security Exchange Commission litigation releases, (ii) extract key attributes of different fraudulent behaviour from unstructured data (iii) create a Business Intelligence system for market surveillance that alerts on possible fraud cases.

Construction of ontology that would support cross-lingual and cross-border financial governance was the goal of the project MONTIFIC (Multilingual ONTology for Internal Financial Control). The ontology aimed at capturing multilingual terminology that would support information resources of the "Internal Financial Control Assessor" training, certification and implementation processes (Budin et al., 2010).

Another multilingual ontology in the financial domain was proposed in project MONNET (Declerck et al., 2010). It addresses the problem of sharing financial reports across European borders, where accounting regulations are different. The use case is based on XBRL reports and XBRL taxonomies. It offers not only a mapping of concepts from different languages, but also the extraction and integration of financial data from reports in different languages, and regulations policies.

European FP6 project DIP developed financial ontology to support e-banking operations scenario, through the use of semantic web services (Montes et al., 2005). Ontology creation was necessary to provide a conceptual map to classify entities and their relations, establishing a common understanding on used terms. The ontology modelled only the mortgage domain. Developing ontology that would cover the whole financial domain was out of the scope of the project, mainly due to the fact that financial market moves very dynamically but standardisation

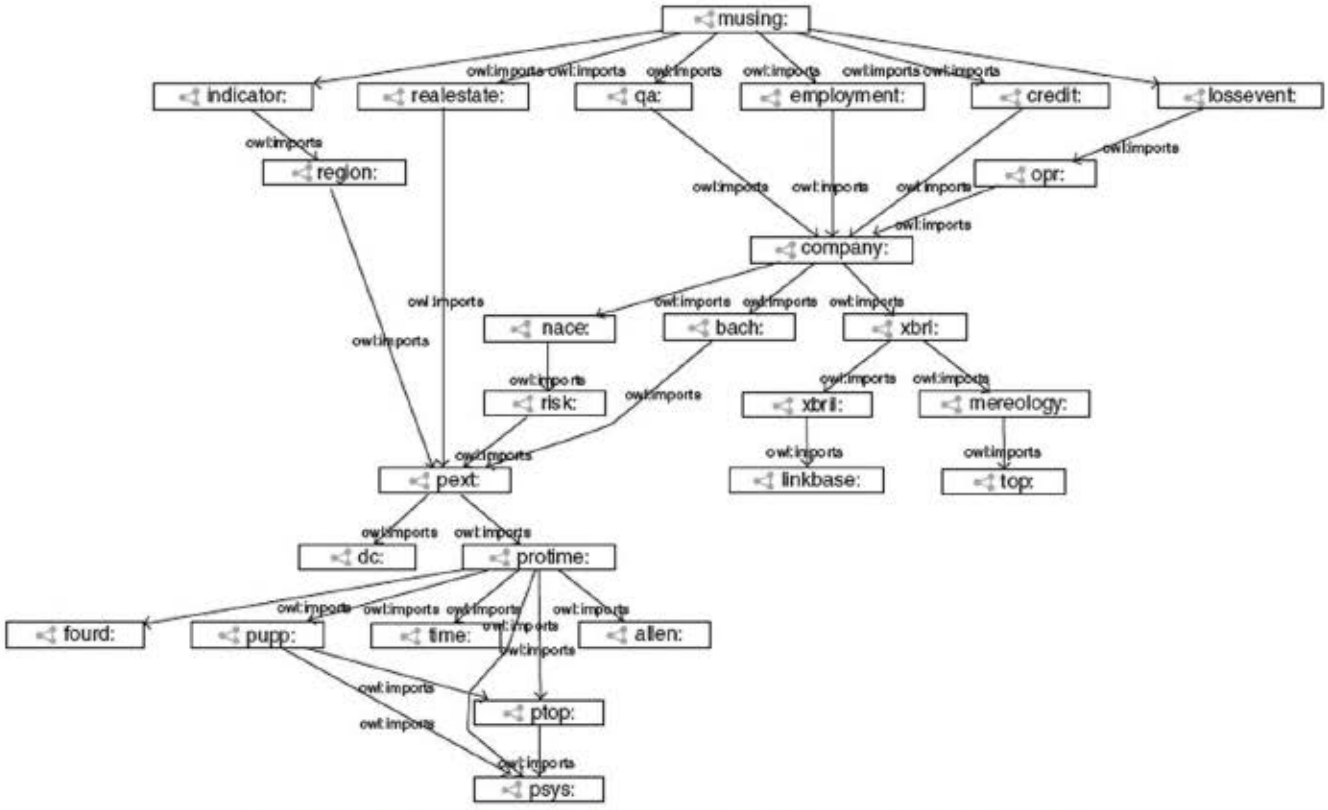



Figure 2.7: Complete import graph of MUSING ontology. The graph shows main components of the ontology and included vocabularies and upper-level ontologies. This structure highlights modularity that allows for ontology reuse in the widest scale possible (Leibold et al., 2010).

efforts are very slow and prone to failure (Hepp, 2007).

Another European research project MUSING provides an example of financial ontology development for Business Intelligence domain for (i) business reporting (Spies, 2010), (ii) operational risk management (Leibold et al., 2010) and (iii) Information Extraction (Saggion et al., 2007). The MUSING ontology buils on top of previously established work in the field, by importing relevant domain-independent ontologies and other standard reference ontologies. Figure 2.7 shows the complete graph of ontology imports, representing a conceptual model of the ontology structure.

Ontology creation is also an important step in the Information Extraction task, by providing a backbone for concept mapping between

entities found in textual sources and structured ontologies. One of such example was presented in the EU FP7 project FIRST. The FIRST ontology was bootstrapped based on gazetteers and other structured sources, and then the iterative ontology evolution process was applied in order to refine and aggregate more relevant concepts. The ontology is oriented at the financial domain for fostering text mining and annotations of such concepts as companies and organisations, markets, stocks and stock exchanges, currencies, countries and geographical locations, important actors in finance and politics, sentiment vocabularies and other indicators for financial instruments. All concepts provide also necessary gazetteers support that indicates ontology annotation specific data, such as company suffixes stop words, stock acronyms, currency symbols and others (Grčar et al., 2012).

As sentiment quickly turned into another important indicator in financial domain, it became a subject for more structured semantic modelling. Existing generic and domain-independent vocabularies for opinion mining (Westerski et al., 2011) were adapted to deal with web-based content (Sánchez-Rada and Iglesias, 2013; Sánchez-Rada, Vulcu, et al., 2014) and applied to financial news sentiment modelling (Sánchez-Rada, Torres, et al., 2014).

Substantial effort has been spent in transforming structured and semi-structured datasets into semantic form for further interlinking with other datasets, improved querying, data publication and other data integration tasks (Auer, Dietzold, et al., 2009; Sahoo et al., 2009). The whole EU FP7 project LATC (The Linked Open Data Around-the-Clock) has been devoted to fostering the creation of LOD datasets from public or otherwise siloed or unavailable datasets (Hausenblas, 2010). However, the sheer idea of semantifying existing data (also called "triplification") by straightforward transformation into RDF triples is not enough and additional semantic modelling is usually required in order to provide more navigable or semantically sound form. Project FLORA was oriented at creating financial knowledge base based on structured data sources, such as XBRL filings (Radzimski, Sanchez-Cervantes, et al., 2014). Many

previous efforts to semantify XBRL data were merely mimicking XBRL structure, thus providing unnecessary burden for adopted business use cases as they still required a deep understanding of XBRL-specific modelling. Therefore a shift from canonical to an Entity-Attribute-Value (EAV) model was adopted in FLORA. This approach, in conjunction with RDF reification, fostered comparative analysis and financial ratio calculation process (Sánchez-Cervantes, 2015). The resulting conceptual model provided more understandable structure and better human-navigable form.

All the previous efforts to construct the financial ontology were mostly oriented at concrete business domains or aimed at modelling a smaller subset of the bigger financial world. Providing a comprehensive model that would cover the whole domain is a difficult task, that needs broader consensus and involvement of different parties. In 2005 Enterprise Data Management Council (EDM Council) was established in order to establish standards for effective data management for the development of business relationships. Having a commitment from many business and financial institutions, EDF Council started development of the Financial Industry Business Ontology (FIBO). The goal of FIBO is to become the open industry "common language" standard for defining the terms, facts and relationships associated with financial contracts (Bennett, 2013). The ontology is divided into the following domains: (i) Foundations, (ii) Business Entities, (iii) Indices & Indicators, (iv) Financial Business & Commerce, (v) Securities & Equities, (vi) Derivatives, (vii) Loans & Mortgages, (viii) Market Data, (xi) Funds, (x) Corporate Actions, (xi) Portfolio/holdings, (xii) Payments. As of 2016 four FIBO standards have been already released as Object Management Group (OMG) standards. Those are: FIBO Foundations[10], FIBO Business Entities[11], FIBO Indices and In-

[10] The current version is at: http://www.omg.org/spec/EDMC-FIBO/FND. Last accessed: 24/04/2017.

[11] http://www.omg.org/spec/EDMC-FIBO/BE. Last accessed: 24/04/2017.

dicators[12] and Financial Business & Commerce ontology[13].

Results of this development shows that FIBO is becoming the reference ontology due to the significant standardisation efforts, substantial support from the industry and high coverage of the whole financial domain. However, at the time of writing of this thesis, FIBO is still in an early stage of development. Only FIBO Foundation has passed version 1.0. The remaining three released ontologies: FIBO Business Entities, FIBO Indices and Indicators and Financial Business & Commerce are still in "beta" versions.

The following thesis employs semantic technologies for representation of knowledge extracted from unstructured data. The Linked Data form will foster the integration of financial data coming from various sources through the interlinking of main concepts. As a consequence that will lead to better-informed financial decisions by blending the knowledge from other knowledge bases, improved knowledge querying and analysis and interoperability with other financial systems.

## 2.3 *Text analytics for decision support in finance*

The origins of Decision Support Systems (DSS) date back to late 1960's when the first studies of computer-aided decision systems were conducted (Ferguson and Jones, 1969). Since the beginning, the DSSs aimed at providing support for decision-making process for solving complex domain-specific problems using decision models (Klein and Methlie, 2009). Especially in areas where the amount of information is too big for human to easily analyse or the nature of the problem is very complex. In this sense, the decision support system makes extensive use of (i) data (ii) documents (iii) knowledge bases (iv) decision models in order to enhance one's ability to make better-informed decisions. There exist many taxonomies that propose to organise Decision Support Systems into coherent groups and to structure the vast majority of examples into cate-

[12] http://www.omg.org/spec/EDMC-FIBO/IND. Last accessed: 24/04/2017.

[13] http://www.omg.org/spec/EDMC-FIBO/FBC. Last accessed: 24/04/2017.

gories based on e.g. generic operations (Alter, 1977), knowledge source (Holsapple and Whinston, 1996) or organisational class (Hackathorn and Keen, 1981). With an increasing number of DSS, shifting of focus and technologies, taxonomies evolve to accommodate to the new scenarios. For instance, Power (2001a) proposes a framework for classifying Decision Support System into five main categories: (i) Data-driven DSS, (ii) Model-driven DSS, (iii) Knowledge-driven DSS, (iv) Document-driven DSS and (v) Communications-driven and Group DSS, and three secondary components: (i) Inter-organisational or Intra-organisational DSS, (ii) Function-specific or General Purpose DSS and (iii) Web-Based DSS.

Decision Support System have been applied in almost all industries and finance quickly became one of the most interesting research areas. The first commercial DSS, called IFPS (Interactive Financial Planning System) was developed in the late 1970's at the University of Texas (Power, 2007). Since then, the diversity of financial Decision Support Systems have flourished. From the standpoint of their application, DSS in finance can be categorised as follows (Alić et al., 2012):

- Financial analysis systems targeting the problem of predicting future movement of individual stocks, stock indices, bond ratings, etc. in order to improve portfolio management and investment decisions.
- Analysis of risk factors and detecting of events that might influence change of future ratings, volatility or other undesirable events that can be summarised as Risk Management.
- Finding potentially fraudulent behaviour in financial operations, detecting false or otherwise doctored financial statements, performing market surveillance and other market abuse monitoring for fraud detection.

Financial Decision Support System are aiding humans to understand and take actions in the ever-changing situation of the world. Therefore, apart from the internal decision models, and knowledge bases, the main component is the source of external data from the markets, economical

situation, transactions, etc. The principal source of data for such systems can be divided in two main groups:

- Structured financial data, such as earnings, quarterly reports (e.g. 10-Q XBRL filings), financial ratios, stock price time series, sales data, macroeconomic indicators, transaction logs
- Unstructured data, such as corporate announcements, news, experts analysis and opinions, other textual sources accompanied by structured documents.

Structured data represent a state of the world of chosen aspects, such as company financial situation, using some agreed methodology and common metrics. It is a way of presenting hard facts using numbers. Such data have mostly well-defined semantics, thus can be directly applied to models and algorithms, compared with other data and has certain interpretation. In consequence, such data is easier to act upon. Structured data is also more trustworthy, it is subject to more scrutiny and legal regulations. Manipulations of such data happen rather rarely.

On the other side, there is unstructured data, an important source of information in finance, used extensively in the decision-making process (see Figure 2.8). Unstructured data, such as news, also present changes to the state of the world through the expression of facts, as well as opinions, rumours, predictions, etc. Such data is not directly actionable in a fully automatic manner but needs to be processed first. It can be either read by a human who decides on the action or processed by automatic NLP system in order to extract relevant features for further decision-making. In either case, unstructured data is less certain: news can be overstated, some sources can spread false rumours and predictions not always accurate. Also, the process of information extraction from such sources might not be fully accurate, lack proper semantics and requiring additional human effort. This makes unstructured data harder to deal with and more challenging.

Decision Support Systems in financial domain that deal with unstructured data aim at overcoming those problems and provide automatic


Figure 2.8: Unstructured data are used on a daily basis in financial decision-making process. Here the Bloomberg terminal is showing live Twitter feeds on selected stock symbols. The service is very popular among traders around the world.

means for information extraction from unstructured data in order to employ them into the decision-making process.

Micu et al. (2009) introduces StockWatcher, a financial news analysis system, employing NLP and text mining techniques and domain knowledge, expressed in a form of OWL ontologies, in order to detect relevant news that can affect investors' portfolio. The knowledge base contain a glossary of keywords for detecting the following financial events in the news: (i) analyst forecasts, (ii) contracts, (iii) earnings, (iv) results, (v) sales, (vi) stocks and shares, (vii) acquisitions, (viii) collaborations, (ix) company performance and (x) new products. The further step is to assess if the news has a positive, negative or neutral effect on the stock price of the company of choice, e.g. if earnings are lower, then the overall effect would be negative. Based on its set of rules, the system further categorises news and gives final recommendations to the investor.

A slightly different system for news analysis has been presented in (Nuij et al., 2014). This work proposes a framework for deriving trading

strategies based on technical indicators extracted from the news. It is based on financial event detection in order to classify the news about business entities. For classification step it is using proprietary system ViewerPro that is using keyword-based methods for assigning the news to one of the following classes: (i) shares up, (ii) rating up, (iii) collaboration start, (iv) rating positive, (v) profit higher than expected, (vi) acquisition start (vii) sales up (viii) price target raised (ix) business expand, (x) joint venture, (xi) performance exceeds expectations, (xii) profit up, (xiii) collaboration consideration, (xiv) performance meets expectations, (xv) profit down, (xvi) rating down, (xvii) price target lowered, (xviii) shares down, (xix) profit lower than expected. What is noteworthy is that the expected effect of each event is already included in the category (e.g. profit up vs. profit down), so no further analysis is necessary. The evaluation is performed by analysing average returns and abnormal returns for each event and creating a trading strategy by incorporating selected extracted indicators and finding optimal results.

Using news articles as a source for constructing Business Intelligence system was studied in (Chung, 2014). It describes the extraction of the relevant BI factors in the process of text mining in order to provide relevant information to business analysis and assist them in the process of decision-making. The definition of factors is guided by experts in the field and the information retrieval is based on the textual features selection and extraction. The categorisation of BI factors was performed using Naïve Bayes algorithm.

Detecting events in the news is an important field studied in the context of the financial decision-making process. Java et al. (2006) presents SemNews system for extracting entities and facts from the news. It uses OntoSem text processing engine (Nirenburg and Raskin, 2001) for information extraction that employs frame-based language for the knowledge base and onomasticon for proper names resolution. Extracted facts are further translated into an OWL ontology, as original OntoSem system uses its own knowledge representation format.

In the context of project PARMENIDES, Black et al. (2005) proposed

CAFETIERE system for automatic entity annotation and relation extraction from the news. It is based on the GATE framework and ANNIE IE pipeline, with custom extraction rules that operate on a lexical and semantic level. It is using hand-crafted rules and gazetteers for annotation and entity lookup.

A different aspect of information extraction from unstructured sources has been proposed in SONAR (Gómez-Berbís et al., 2009). The proposed architecture uses NLP techniques to extract certain facts from news and populate the ontology in order to perform semantic search and reasoning over the internal knowledge base. The Analyst Information Assistant Module (ANNE) built on top of SONAR aims at guiding analysts in investment planning and financial decision support.

Even more complex approach to event extraction is presented in (A. Hogenboom, F. Hogenboom, Frasincar, et al., 2013). Authors introduce Semantics-Based Pipeline for Economic Event Detection (SPEED), that aims at detecting and extracting relevant financial events from news and automatic annotation with the use of ontologies.

Another area where analysis of unstructured data has been studied is automated extraction of sentiment, called sentiment analysis or opinion mining (Pang and L. Lee, 2008). The main motivation behind this is that human decisions are influenced not only by information but also, and to a great degree, by emotions (Kahneman and Tversky, 1979). The irrational aspect of the human decision-making process is in contrast with the efficient market hypothesis. This gave birth to behavioural finance and further studies on how mood and emotions are influencing financial decisions.

There have been numerous studies focused on the applicability of sentiment analysis in finance, and how mood is shaping the stock markets (Brown and Cliff, 2005). Most of them are oriented on predicting: (i) future stock prices movements, (ii) volume (iii) volatility or (iv) investment risk based on sentiment and volume of mentions in social media (Baker and Wurgler, 2007).

While all the work is mostly oriented on verifying hypothesis if there

is any significant correlation between sentiment in social media or other textual sources and stock market trends, they differ in approach, techniques employed and indicators they try to correlate (Nassirtoussi et al., 2014).

Twitter became a social media of choice for sentiment analysis due to the fact that it is relatively easy to obtain big amounts of data on chosen subjects. The Twitter API allows for streaming content of tweets in real-time fashion in order to get timely information for user-defined queries (Pak and Paroubek, 2010).

Using twitter for sentiment analysis in order to derive a trading strategy was studied in (Bollen et al., 2011). The dependence between Twitter sentiment and the Dow Jones Industrial Average (DJIA) index was observed in (Ranco et al., 2015). The study found significant Granger correlation during the peaks of Twitter volume. In (Nann et al., 2013) authors achieve better performance than the S&P500 by including twitter sentiment in their trading model. However Nofer and Hinz (2015) observe that a simple mood aggregation from Twitter is not enough to show predictive power on stock markets but additional analysis of the social structure is necessary. This could be also due to the fact that more investors are already including Twitter in their trading strategies. Markets are adapting to that fact and become more efficient, diminishing the advantage of sentiment trading.

While Twitter is a popular source of unstructured data, other more conventional data sources are also widely used for mood analysis and extraction of other relevant features. Most used sources are news articles, bulletin boards and blog posts, public companies' disclosures and corporate financial reports and others. Typically acquired through RSS[14] feeds in the process of web crawling, they can also provide timely information that can be used to make financial decisions. (Gilbert and Karahalios, 2010) analyses the mood of blog posts from LiveJournal and discovers that it has a significant influence on the S&P500 index. Analysis of public

[14] RDF Site Summary, Rich Site Summary or later called Really Simple Syndication — a format for publishing frequently updated information, such as news or blog posts.

disclosures for making short-term investments was presented in (H. Lee et al., 2014). The work focused on 8-K forms published on the Securities and Exchange Commission (SEC) website. Those forms provide information on the major events of public companies and authors demonstrate that information from these reports influences the future stock price of the corresponding company.

Goonatilake and Herath (2007) analysed news in order to build a regression model on volatility. In (Groth and Muntermann, 2011) authors proposes an analysis of textual sources for financial risk management. Analysis of corporate disclosures and extraction of valuable information confirmed that unstructured data can improve detection of supranormal risk exposures. In (Leinweber and Sisk, 2011) authors are analysing news in order to derive signals important for making investments. They construct a model for portfolio construction and obtained good performance, with annualised Sharpe ratio of 0,76.

## 2.4 *Conclusions*

As we have observed, there is a broad number of examples of work concerning unstructured data analytics and particularly employed in the domain of finance. Ranging from event detection, sentiment analysis or Business Information Systems and ending in extracted knowledge representation techniques, each of them targets a very concrete and narrow field of interest. While previous work described here is highly relevant and cover similar domain, we find many shortcomings that this work aims to address. In this sense, the presented work is different from previous approaches in several aspects:

1. It establishes the overall process of Information Extraction, from data acquisition to Decision Support, covering the whole data life cycle and semantic knowledge representation. Not only does it cover the information extraction phase, but also aims at providing sound BI market insight platform based on extracted knowledge and backed with DSS.

2. While others are successfully employing semantics in the process of financial decision-making, they largely differ in the applications that range from ontology-based Named Entity Recognition to create ontologies out of extracted data. In this work, we are rather modelling facts and indicators that constitute a knowledge base for the decision-making process. The dataset with extracted financial facts is the biggest to our knowledge covering the Spanish market. The resulting semantic dataset is compatible with the Linked Data approach, available for querying, interlinking and augmenting with other datasets, thus increasing the added value of extracted knowledge.

3. The outbreak of deep learning techniques and their recent application in Natural Language Processing provide new means for Information Extraction. This largely improves the process of information extraction and decision-making comparing to aforementioned previous approaches. We aim to minimise the human effort and use machine learning techniques rather than hand-crafted rule-based systems.

4. The previous work on news analytics is mostly focused on word-based classification to extract relevant information. Such approach has obvious limitation by largely missing word semantic and results in less accuracy. This work employs semantic word vectors in the NLP phase in order to improve the phase of information extraction.

5. The context of Spanish market implies the development of several corpora for text classification experiments. There are relatively few language resources in Spanish for this concrete domain. As such linguistic datasets were not available before conducting this work, we believe those are valuable assets as well. We plan to release those artefacts to the community in order to build on top of our work.

It is also worth noting the limitations of this thesis. Although we are working with unstructured, noisy and sometimes less reliable data, we are not planning to deal with the uncertainty of news, false rumours or hoaxes. We rather aim at carefully selecting data sources in order to avoid and minimise the risk of noisy data.

# 3 Research hypotheses & methodological evaluation approach

As we observed in previous chapters, the state-of-the-art highlights promising research directions for BI systems in the financial domain. This combined with previous experience, mainly based on already mentioned projects FLORA and FIRST establishes a sound base for carrying out further work within the lines established. Combining data-based experimentation, recent advancement in the field of Information Extraction and NLP and results from previous projects we propose a set of hypotheses that will be verified within this work.

The main objective that this thesis aims at resolving is the viability of incorporating unstructured data into financial decision-making and provide meaningful insights into the constantly changing world of finance based on rapidly flowing stream of textual information. The work is carried within the domain of finance by employing text analytics and semantic modelling to enhance the decision-making process.

The following sections aim at providing a structured guidance into the rest of this work by breaking down the objective of this thesis into several sub-hypotheses and establishes the process of evaluation leading to verification of those hypotheses. The high-level objective is refined into concrete sub-hypotheses, and relation with objectives is presented in the next sections.

## 3.1 Hypotheses

The central objective of this thesis is split into four main hypotheses:

*H1:* Automated ontology-based information extraction of unstructured financial data leads to better decision-making process in the financial domain, and Business Intelligence in particular.

*H2:* Encoding semantic facts in the process of machine learning improves text classification:

- Knowledge-based systems help improving the process of extracting information from unstructured financial data,
- Semantic technologies improve accuracy in mining unstructured financial data.

*H3:* Semantic knowledge base created from unstructured sources improves analytical capabilities in the financial domain.

*H4:* The use of deep learning techniques in information extraction can improve the process of text analysis in the financial domain as compared to the traditional machine learning methods.

Note that the hypotheses are tested in the framework of Decision Support Systems as a realisation of the decision-making process in the context of Business Intelligence, as a particularisation of financial-domain.

## 3.2 Investigation process

The investigation is driven by data transformation process and comprises all steps necessary for turning unstructured data into relevant information for financial decision-making. Figure 3.1 presents the high-level overview of the process that forms the analytical pipeline and its main stages, starting from acquisition of raw data, through information extraction steps and finishing in evaluation of the decision model.


Figure 3.1: Stages of the investigation process

Each stage has the following role in the overall process:

1. Data acquisition: The first stage aims at acquiring unstructured data that are relevant for the decision-making process, and establish continuous data acquisition process, where such data can be retrieved on a regular basis. Primary task is to analyse sources of unstructured data, such as: financial news coverage, corporate disclosures, market reports and analyses, expert and non-expert opinions on blogs and websites. Their feasibility for decision-making is assessed through the literature and experts' opinions in order to filter most relevant sources.

2. Feature analysis: In order to focus only on relevant, we are performing an analysis of most promising features to extract from the data. Based on domain knowledge, previous work and data analysis we choose a set of high-level features to be extracted from text. Those are the features that indicate relevant financial events and are important from the point of view of risk analysis and decision-making process.

3. Semantic modelling: This stage provides mostly conceptual work with the aim of modelling extraction results in a form of semantic graph. Based on the feature definitions to be extracted from unstructured text, we model the knowledge in terms of semantic concepts (classes) and relations (properties). Semantic knowledge representation provide various advantages in the area of data interoperability, sharing and publishing and knowledge inference.

4. Information Extraction: A this stage, the natural language processing pipeline is analysing documents from the web and extract relevant features that can be used for the financial decision support. The Information Extraction process comprise multiple steps, each one special-

ized in extraction of the concrete feature set. It starts with performing lower-level tasks, such as text cleaning, segmentation, tokenisation and ends with named entity tagging, relation extraction and text classification tasks.

5. Decision model creation: The high-level features extracted in the previous step establish solid foundations for developing the model for financial decision-making process. The semantic knowledge base is an input for the decision support model training. It is an iterative task in which the data driven evaluation is necessary in order to assess the viability and quality of the overall result. The whole process is data-driven, based on the real market data correlated in time with the financial events.

6. Evaluation: The decision support model takes into account different categories of financial events to measure the impact of detected events for the particular investment, as learned using the historical data. The evaluation aims at deciding the performance of the model, its utility and viability in the use-case scenario.

The order of all stages is sequential, but each single part is developed iteratively rather than sequentially. Therefore it does not necessarily imply the waterfall-like process, where the task sequence, duration and dependencies are defined in a more strict fashion. Rather we have adopted a more iterative methodology, where sometimes a refinement of some previous task or some "cross-task" work is necessary.

The motivation behind this is to enable experimentation and prototyping, where the most important factor is validation and verification of main hypotheses. Also due to the fact that some stages are interdependent of each other, and a change in one imply changes in another. This approach also enables to "close the loop" earlier, even if previous stages need refinement.

In the following chapters we will describe in details the overall process and its stages in the following order: Data Acquisition and Corpus Creation will be described in Chapter 4, Feature Analysis and Information

Extraction in Chapter 6, Semantic Modelling in Chapter 5 and Decision model creation and Validation in Chapter 7. Summary and Conclusion will be presented in Chapter 9.1.

# 4 *Data acquisition for the analytical pipeline*

This chapter focuses on the process of news acquisition in general which is the first step in the decision making process. We are describing the first stage of the investigation process called "Data Acquisition", as defined in the methodological approach (see Section 3.2). Gathering the unstructured data is the most fundamental dependency and a prerequisite for next steps.

Decision-making in the financial domain is largely driven by information. All actors act upon what is known to them and what they learn from external sources. When it is a rational, well-informed investor or irrational, so called "noise-trader" they are all making decisions based on some chunks of information that they receive and consider relevant. The variety of such information sources may be very broad: from social media, analysts' opinions, official media agencies, insider information[1], global economic indicators, rumours, official corporate statements, and many more. At the end of the day, most of such information is turned into news and published for broader audience through various media channels.

News are crucial to the financial markets. They report events that might of a interest to investors or general public. By spreading the information they influence the market to the great extend. "News moves the market" is a popular phrase, but the growing body of research confirm that news can influence the asset price, volatility and assets' risk (Da et al., 2011; Dzielinski et al., 2012). The rapid access to the news feed is therefore important for people making financial decisions such

[1] In most cases trading on insider information is illegal.

as traders, investors, portfolio managers, etc.

From this point of view, a news can be seen as a carrier of events. An event is something that happens at some point of time that changes of the state of the world and might have important consequences in the future. Such event may affect the value of company stocks, securities or influence the risk associated with them. News that describe such event is sometimes called "material news". Other events might have no direct influence in the short term, but still be relevant in the long-term. Some of them might be very simple and other can be of a very complex nature, involving many actors and provoking a chain of subsequent events.

The amount of available news is growing constantly, but most relevant events are rather few. Even fewer are potentially disastrous events or other unexpected occurrences that need immediate reaction. Therefore, it is important to provide enough coverage to capture events of different kind even if they are buried in the big stream of noisy or otherwise less relevant data. The context of Spanish market is also an important aspect. While it is easier to achieve good coverage for English-language media by capturing news on Dow Jones Industrial Average (DJIA), Standard&Poor's (S&P500) or Financial Times Stock Exchange (FTSE), more effort is needed for the Spanish market that comprise far fewer public companies.

The following sections describe the general news source selection, data acquisition process, preprocessing, corpus preparation and corpus evaluation.

## *4.1 Data source selection*

Acquiring unstructured data that are relevant for the Business Intelligence and financial decision-making process is the first stage of the text analytics task. This involves selecting data sources and establishing continuous data acquisition task where data can be retrieved on a regular basis. As pointed out in the goals of this thesis, we focus on longer texts and our primary sources are textual news and official company announcements.

Most news outlets and services provide news for a broad range of topics, such as national politics, foreign affairs, economy, technology, sports, science, culture, lifestyle and columnists' opinions. There are also specialised sources, that aim at a concrete topic such as finance or sports. For the sake of this thesis, we compiled the list of most influential news services that provide accessible and crawlable online version of the published news. We aimed at mass audience news sources but also online newspapers that target domains of business and economy.

The choice of news sources used for the corpus creation was based on popularity of the sources and availability of news archives for crawling, as explained in 4.2. We focused primarily on financial news sources, and then on general news sites, covering also economy and finance stories. Appendix A gives a quantitative overview of the news sources that were selected for crawling.

The second data source are corporate disclosures of public companies that are published by every company issuing stocks or other financial instruments. Such public filings (called *hechos relevantes*) must be published every time when a significant business event occurs[2]. Such events might affect the value of financial instruments, associated risk, future business operations, etc. *Hechos relevantes* are publicly available in the database of Spanish national financial market regulator - Comisión Nacional del Mercado de Valores (CNMV[3]).

## 4.2 *Data acquisition*

The data acquisition task was performed in two stages. The primary data acquisition task has been performed continuously during 6 months, from 20 May 2014 until 18 December 2014. It resulted in total 341.849 news documents acquired from RSS feeds of 43 Spanish online newspapers and news services. The first dataset was acquired in order to assess

[2] As required by law: Artículo 82 de la Ley 24/1988, de 28 de Julio, del Mercado de Valores and latest changes in Artículo 228 de Real Decreto Legislativo 4/2015, de 23 de Octubre, por el que se aprueba el texto refundido de la Ley del Mercado de Valores.

[3] http://www.cnmv.es. Last accessed: 24/04/2017.

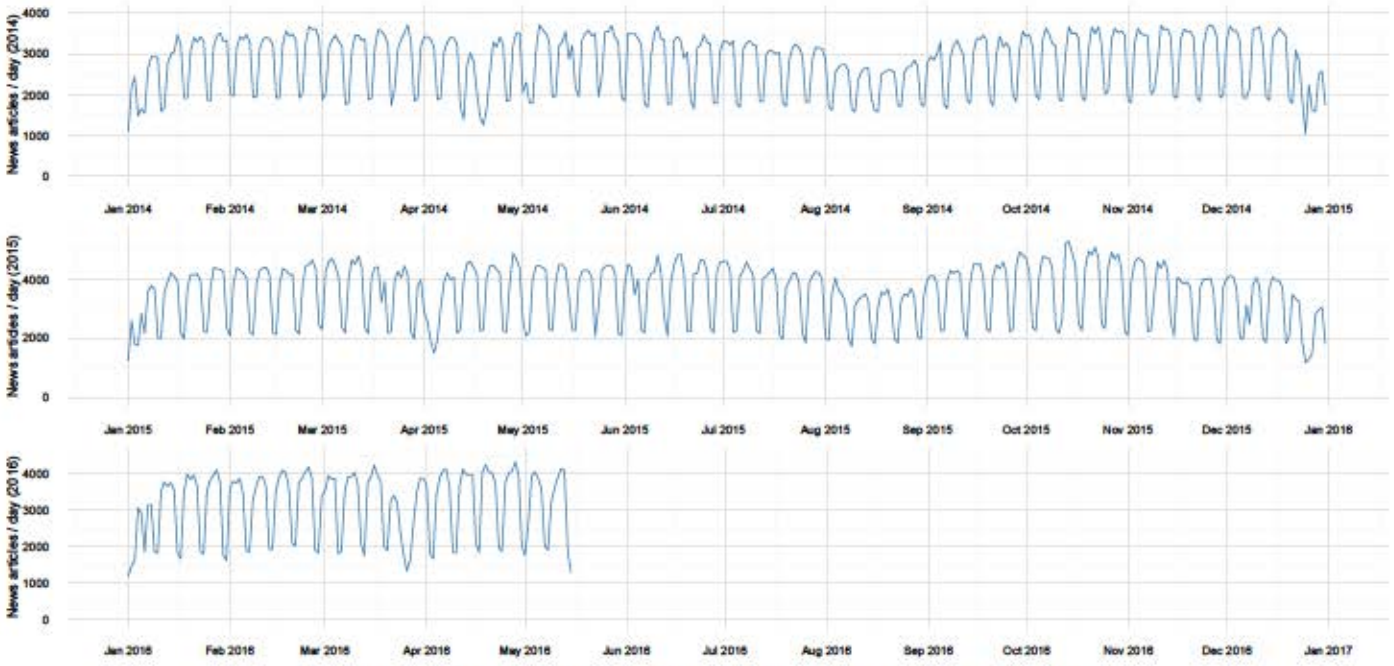



Figure 4.1: Volume of acquired raw news documents per day. Visible tall spikes represent daily peaks on weekdays. This chart shows the final volume after all cleaning and normalisation steps.

the viability of the continuous process of news extraction.

In the second iteration, the crawling was extended in order to get bigger news coverage, extending beyond the initial 6 months of 2014. Also an effort was put to better balance the complete dataset and normalise any occurring anomalies. This process was performed for the period of 1 January 2014 until mid-May 2016 in order to capture more events of interest. It was done in part retroactively, by crawling through news archives (*hemerotecas*).

The complete news dataset resulted in the total of over 2,8 million documents, after normalisation. Further subsections give an overview of normalisation steps and dataset anomalies that needed to be accounted for.

Overview of the daily volume of raw news acquired in the crawling process is presented in Figure 4.1. Note that the periodicity observed in the graph resembles daily and weekly patterns of news volume, with more news produced during the weekdays from Monday to Friday and

less during the weekend and bank holidays. There is also significant seasonality during the year. The lower volume in month of August is due to very common holidays season in Spain and the end of December due to Christmas holidays. Other bank holidays (e.g. Easter) are also visible.

In a similar way the complete dataset of CNMV corporate disclosures was harvested. Crawling all public fillings posted from 1 January 2014 until 1 June 2016 resulted in total of 32.564 documents. Table 4.1 provides an overview of both datasets size. It is important to note that directly crawling selected news sources resulted in better more complete dataset, rather than generic RSS crawling.

| | First Iteration<br>May 2014 – Dec 2014 | Second Iteration<br>Jan 2014 – May 2016 |
|---|---|---|
| News | 341.849 documents<br>43 datasources<br>(RSS sources) | 2.870.870 documents<br>25 datasources<br>(direct crawling) |
| Company disclosures (CNMV) | 18.126 documents | 32.564 documents |

Table 4.1: Overview of the news and company disclosures datasets. Dates show the document coverage.

For crawling the news sources we used Heritrix web crawler[4] for the first iteration and continuous crawling engine and the Scrapy python framework[5] for crawling news archives (*hemerotecas*).

[4] Heritrix project: https://webarchive.jira.com/wiki/display/Heritrix. Last accessed: 24/04/2017.

[5] The Srappy source code and documentation can be found at https://github.com/scrapy/scrapy. Last accessed: 24/04/2017.

## 4.3 *Data pre-processing*

Corpus text preparation comprise multiple steps in order to transform downloaded news web pages into clean text that can be indexed and further processed:

- Text pre-processing: extracting relevant meta data about the documents (creation date, news source, documents URL, possible actualisation date, document type).
- Boilerplate removal and text cleaning.
- Noise removal and normalisation, removing any noisy content, such as: ads, erroneously crawled sources and non-textual documents (e.g. videos).
- Removing possible duplicates.
- Assessing the dataset quality.

Note that aforementioned steps apply only to the web corpus dataset. The CNMV dataset consists of mainly PDF files. The text extraction is a straightforward process. We used `pdftotext` Unix tool to extract clean text and further steps were not necessary.

### 4.3.1 *Meta data extraction*

Meta data is additional information about the news document that describe the content and is crucial for proper data indexing, as it contain information that would be otherwise difficult to obtain from the news text only.

After acquiring each document we extract relevant meta data that is immediately available from HTML header, such as news source, document URL and document title. For HTML documents that provide a standard way of encoding date within the `<meta>` tags in the header, we use it as the preferred way of getting document date. However, in practice this is true only for a limited number of documents. In this case we

need to manually define a date location within the DOM structure for a different class of documents. Also the date formats can vary from site to site. Manually defined date formatters are needed for each news site and even for each news category, due to the fact that different sections of the same news website may use different HTML template. Locating dates within the DOM structure is performed using CSS selectors. It is important to provide accurate date information for two main reasons: (i) document without date is of no use from the event detection point of view, (ii) wrong dates might induce errors in the next stages of processing.

After detecting the date string it is important to properly parse the date with all information provided. Standard ISO date formats account only for a few percent of all date formats and it is important to provide manual hints for proper date parsing.

For this task we use templates based on the URI patterns of the news websites. Such template contain three parts: (i) the authority part of the URI to distinguish between different class of documents (ii) possible date locations within the document class (iii) a set of acceptable date formats for parsing for this document class.

Other metadata that is extracted from the document are based on what is provided in the HTML header section, e.g. in a form of Dublin Core annotations (The Dublin Core Metadata Initiative, 2010). Such data can vary from site to site, but in most cases it consists of: (i) document type (e.g. article, video, blog), (ii) author name, (iii) publisher name, (iv) photo associated with the news (thumbnail), (v) short article description, (vi) tags and keywords, (vii) language, (viii) External sites integrations (e.g. Facebook, Twitter).

### 4.3.2 *Boilerplate removal*

The process of boilerplate removal is extraction of meaningful text from web documents, stripping it from unnecessary HTML code and other non-content parts. For example, a typical web page such as a news article is a text file that consists of HTML code necessary for proper dis-

playing the web page to the user. This code contains markup language containing layout information, page elements, CSS styles, JavaScript code and other information, dedicated mostly to functional and visual aspects of the web page.

Also the textual content of such page is normally full of elements that are not the actual news text, i.e. menu elements, links to other parts of the website, short leads of other relevant news, advertisements, publisher's disclaimers and terms of use, comments just to name a few. Figure 4.2 shows visual elements that from meaningful content among other undesired parts.

After taking out the non-content parts, the actual text of the page is only a small fragment of the whole web page and even smaller part of the whole web HTML file. The process of boilerplate removal means therefore not only automatically detecting displayed text within the HTML markup soup, but also telling the actual article text from other undesired text on the page. While it is possible to manually define a text extraction patterns for different websites using e.g. CSS selectors or XPath, it is a tedious task and still prone to many errors. Even websites of the same publisher can have different templates for each news section, moreover, page templates may change over time, making the manual extraction even more laborious. There are many approaches for automatic boilerplate removal (Laender et al., 2002; Pasternack and Roth, 2009). As boilerplate removal provide only means to achieve the objectives of this thesis, we have not researched this topic, but rather rely on the established work in this field. For practical reasons we focus on available, off-the-shelf implementations. E.g. Kohlschütter et al. (2010) proposed the Boilerpipe[6], a very popular boilerplate removal tool, bundled with Apache Tika project. Another quite popular boilerplate removal tool is jusText[7] (Pomikálek, 2011).

For corpus boilerplate removal, the jusText project was chosen due to the fact that it allows for fine tuning its internal heuristics parameters,

[6] https://github.com/kohlschutter/boilerpipe. Last accessed: 24/04/2017

[7] http://corpus.tools/wiki/Justext. Last accessed: 24/04/2017.

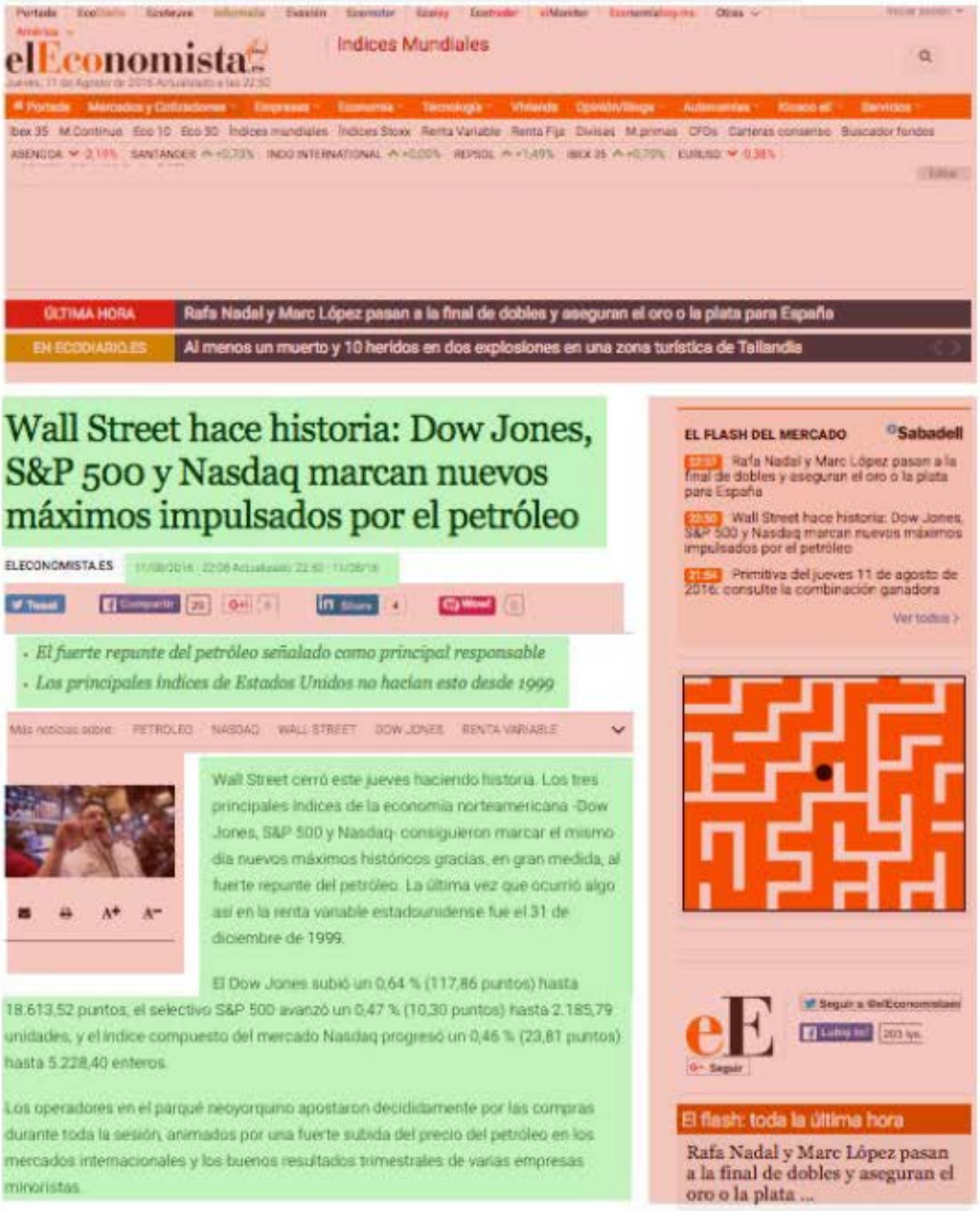



Figure 4.2: Fragment of a webpage, showing content and boilerplate elements. The meaningful text is highlighted in green. Elements in red are undesired.

thus allowing running customised text cleaning for each website class. Figure 4.3 shows an example analysis of a web page with the boilerplate removal text. After adjusting parameters, the same analysis is then run automatically for each document. In the end of this process, we obtain the clean text of the document that is suitable for further processing.

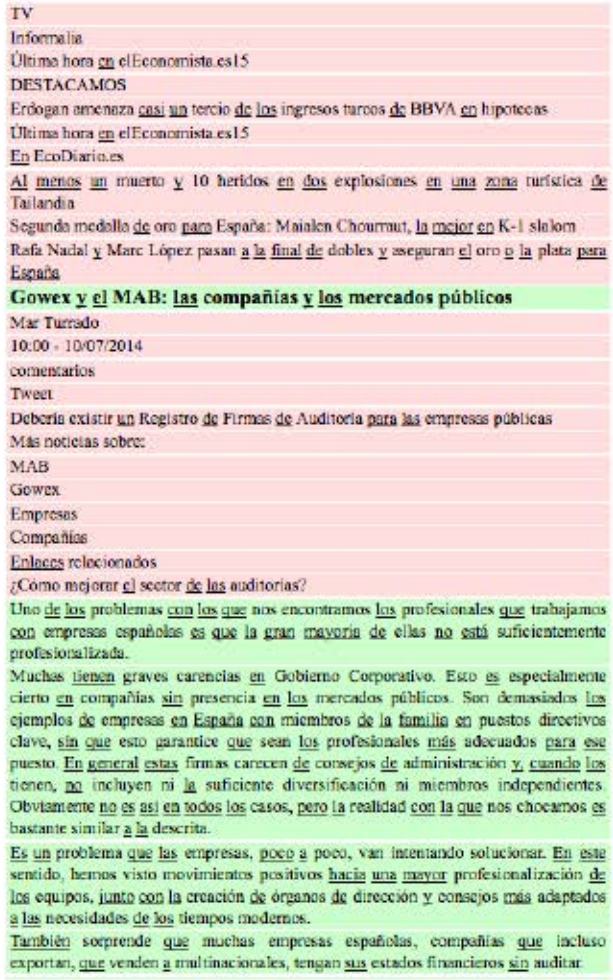

TV
Informalia
Última hora en elEconomista.es15
DESTACAMOS
Erdogan amenaza casi un tercio de los ingresos turcos de BBVA en hipotecas
Última hora en elEconomista.es15
En EcoDiario.es
Al menos un muerto y 10 heridos en dos explosiones en una zona turística de Tailandia
Segunda medalla de oro para España: Maialen Chourraut, la mejor en K-1 slalom
Rafa Nadal y Marc López pasan a la final de dobles y aseguran el oro o la plata para España
**Gowex y el MAB: las compañías y los mercados públicos**
Mar Turrado
10:00 - 10/07/2014
comentarios
Tweet
Debería existir un Registro de Firmas de Auditoría para las empresas públicas
Más noticias sobre:
MAB
Gowex
Empresas
Compañías
Enlaces relacionados
¿Cómo mejorar el sector de las auditorías?
Uno de los problemas con los que nos encontramos los profesionales que trabajamos con empresas españolas es que la gran mayoría de ellas no está suficientemente profesionalizada.
Muchas tienen graves carencias en Gobierno Corporativo. Esto es especialmente cierto en compañías sin presencia en los mercados públicos. Son demasiados los ejemplos de empresas en España con miembros de la familia en puestos directivos clave, sin que esto garantice que sean los profesionales más adecuados para ese puesto. En general estas firmas carecen de consejos de administración y, cuando los tienen, no incluyen ni la suficiente diversificación ni miembros independientes. Obviamente no es así en todos los casos, pero la realidad con la que nos chocamos es bastante similar a la descrita.
Es un problema que las empresas, poco a poco, van intentando solucionar. En este sentido, hemos visto movimientos positivos hacia una mayor profesionalización de los equipos, junto con la creación de órganos de dirección y consejos más adaptados a las necesidades de los tiempos modernos.
También sorprende que muchas empresas españolas, compañías que incluso exportan, que venden a multinacionales, tengan sus estados financieros sin auditar

Figure 4.3: Analysing text blocks of a web page and identifying relevant content vs. the boilerplate text. The text in bold has been identified as title of the article.

The example of an input (HTML) and output of the boilerplate removal (clean text) has been presented in Figure 4.4.

### 4.3.3 *Noise removal and normalisation steps*

Using automatic process for clean text extraction usually has some margin of error. Even adjusting the boilerplate removal tool might result in some percentage of incorrectly extracted text. Such errors can be divided into two groups: (i) when the actual content is missing (ii) when some noisy boilerplate text has been included. The first case, we need to go back to the previous step and ensure proper parameters adjustment

El Banco Central Europeo (BCE) anunciará el jueves, según las fuentes consultadas, una bajada del precio oficial del dinero hasta el mínimo histórico y se convertirá en el primero de los grandes bancos centrales del mundo en cobrar a las entidades financieras por dejar su dinero en la ventanilla de Fráncfort. El jefe del Eurobanco, Mario Draghi, avanzó en mayo que no aplazará más la activación de un paquete de medidas, tras más de medio año de vanas promesas y con un panorama cada vez más japonés: la inflación ha caído al 0,7%, lejos del objetivo del 2% ; el crédito sigue cayendo, y el euro está fortísimo y complica una salida de la crisis decepcionante y lenta. Con esos riesgos en el horizonte, y como complemento del recorte de tipos –que tiene un efecto muy limitado tan cerca ya del 0%–, el BCE tiene listas también nuevas medidas de liquidez para la banca, vinculadas a que las entidades abran de una vez el grifo del crédito, según las fuentes consultadas en Fráncfort y en dos de los grandes bancos centrales nacionales.
Dejará para más adelante la compra de activos al estilo de la Reserva Federal
Después de pregonar centenares de veces que está listo para actuar, a Draghi apenas le quedan cartuchos verbales en la recámara. Su credibilidad está en juego. Debe convencer a los más escépticos de que,

Figure 4.4: Boilerplate HTML code (left) and extracted clean text (right).

in order to capture the main text. In the second case it is possible to perform post-processing adjustments.

In our case, we observed that sometimes extra content is present, such as repetitive copyright notices, cookie policies or related news. Such content has passed through the boilerplate removal and introduces undesirable noise. The removal process is semi-automatic and consists of randomly comparing documents for the same publisher and finding frequently occurring lines of text. Most frequent lines denote a boilerplate content that appears across many different articles and is a candidate for filtering. The only human assistance is needed in order to establish a threshold for repeating frequency that might differ from publisher to publisher and supervise phrases to be removed.

Another step is to verify the soundness of the overall corpus and account for any induced by errors in original data or other unexpected anomalies. In this respect, we detected only two anomalies associated with the wrong original content metadata. In both cases, it is relatively easy to spot such anomalies due to their outlier nature in the aggregated corpus view. Figure 4.5 shows an example of such anomaly. In the concrete case presented here, the huge spike was a result of a bulk upload of video content having the same publication date. The normalisation step was to filter out those documents due to the erroneous dates. As we are not focused on video content, this step does not influence text corpus.

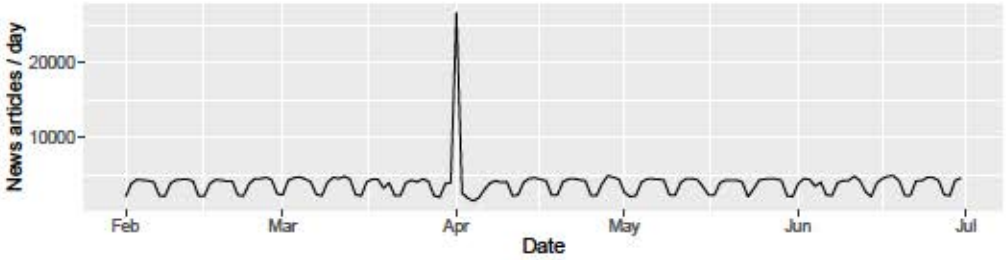



Figure 4.5: Spotting corpus anomalies resulting from errors in original metadata. The aggregated number of documents per day shows an unexpected outlier caused by news publisher error.

Other normalisation steps include proper handling of document encoding, to ensure proper representation of Spanish diacritic characters. All documents has been converted to UTF-8.

The next normalisation step is the removal of duplicate documents. Due to the fact that we are using URLs as unique document identifier, we treat it as sufficient de-duplication technique. There is still possibility of having duplicate documents, as some news are coming from the same news agencies (e.g. EFE). Many financial events will be anyway reported by multiple news sources, therefore this aspect is not going to be addressed at this stage.

## 4.4 *Corpus evaluation*

Corpus evaluation is an important task in order to assess the quality of the news data that will be used as the input in next steps. Quantification of different characteristics allows the assessment of the extraction of text from the web. The mostly automatic process naturally introduce errors in the extracted text. Whether it is an incomplete extraction of chosen features or inclusion of undesired boilerplate text, it all defines the quality of the textual corpus.

While human evaluation of the entire corpus is not feasible, therefore we select a representative subset of documents where news sources are covered in a proportional way. With such constructed sample we perform a manual review, considering the following criteria in the order

presented:

1. The correctness of metadata
2. The completeness of text extraction
3. The amount of boilerplate in the extracted text

Note that when a document fails on one criterion, the further criteria are not evaluated. The evaluation is performed based on the randomly selected sample of 200 documents. Table 4.2 presents results of the evaluation.

| | Percentage |
|---|---|
| Wrong date or date not extracted | 2,5% |
| Wrong title or title not extracted | 0% |
| Wrong character encoding | 0% |
| Content not extracted | 0% |
| Text extracted but not complete | 1,5% |
| Significant amount of noise present | 5% |
| Text with some minor boilerplate present | 26% |
| All text correctly extracted | 65% |
| **Total admissible documents** | **91%** |

Table 4.2: Evaluation of corpus with regard to the clean text and metadata extraction. The upper section shows documents that are considered inadmissible, while the lower one — the suitable ones.

The results are divided into two groups: the first group consists of documents with serious flaws that exclude them from further processing, such as the lack of date, or a significant amount of boilerplate. In the case of the latter, this may mean that some undesired content has been included, such as user-generated comments to the article, that can significantly influence further processing. Such documents can be treated as noisy (e.g. with a bigger amount of boilerplate or comments present) or not admissible at all (e.g. no date present). On the other hand, the second group consists of documents where the clean text has been properly

extracted or some minor boilerplate is present. By minor boilerplate we mean a single word or phrase that is not a part of an article (e.g. *"¿Compartes?"*, *"Regístrese"*, *"Suscríbete al boletín"*) but has been included in the clean text. Typically, the presence of a minor boilerplate is at the beginning or the end of an article. In all seen cases the boilerplate is a static text (that is, a set of repeating words or phrases) and is common for the same news source. This makes it very easy to filter out as opposed to the cases where a significant amount of boilerplate is found, which happens to vary across documents.

The result of 91% of the clean text is a number good enough for using it as a source for further analysis. While in the subsequent analyses we can rely on such input, it is important to note, that this corpus itself can be further improved in terms of clean-text quality. The process of the corpus curation is an iterative process, and in the future, the quality of the clean-text extraction can be further improved e.g. by applying other more advanced approaches. This task, however, is beyond the scope of this thesis.

## 4.5 *Conclusions*

The "Data Acquisition" is the first step towards a decision-making process. The identification of input data sources that are relevant in the decision-making process is the first and foremost prerequisite for any data-driven Decision Support System. In our case, we deal with unstructured data that, contrary to structured data, lacks any organisation and require extra pre-processing steps in order to transform it into a usable input data for further processing. Especially important in the case of news text extraction is the boilerplate removal step. It is precisely there where the human readable part is separated from the mark-up language boilerplate. We showed that although it is not a straightforward task, it is reasonable to approach it in a fully automatic fashion. We obtained around 91% of admissible documents, by applying automatic, heuristic-based approach (Pomikálek, 2011).

Another important observation is that even at the very first stage we

already have to deal with real-world problems such as noisy sources, data acquisition anomalies, pre-processing errors, content duplication and so forth. It is, therefore, crucial to deal with those issues in a proper way to avoid any biases and other errors that could influence future results.

This step is also important from the Information Extraction point of view. While we are not yet applying any NLP techniques, we have already extracted some document metadata that are the first features (*low-level* ones) to be used in the future knowledge base creation. So far we index all data in a purely "syntactic" way, however, this representation will also be transformed into semantic triples.

# 5 *Semantic modelling of high-level features*

After learning a piece of information we would like to know what it is about and how to classify it according to our current knowledge. We need to assign a *meaning* to it, so we can later refer to that bit of knowledge with its semantic properties rather than syntactic ones. The aim is to move from human-oriented, textual event descriptions into more formal, unequivocal and unambiguous representation that can be later processed by machine. Using semantic knowledge representation further allows us to benefit from the entire Semantic Web stack. This includes features such as semantic reasoning, querying, data integration by interlinking with other Linked Data datasets and more.

This chapter focuses on semantic modelling of financial events. The resulting model will further be used to represent financial facts extracted from the news texts described in Chapter 4. The resulting dataset of semantic statements will become a "common language" for expressing knowledge for the Decision Support System. This knowledge will consist of what we call *high-level features*, as opposed to lower level syntactic features. The high-level features are further becoming an input to the DSS in order to support the decision-making process.

## 5.1 *Foundations of the semantic model development*

This stage provides mostly conceptual work with the aim of modelling extraction results in a form of semantic graph. Based on the feature definitions to be extracted from the unstructured text, we model the knowledge in terms of semantic concepts (classes) and relations (properties). Semantic knowledge representation provides three main advantages:

- Provides interoperability through reuse of established ontologies and vocabularies, common understanding of terms and concepts, and unequivocal representation of knowledge.
- Allows for publishing data in a form of Linked Data (Radzimski and Sánchez-Cervantes, 2012) and linking to equivalent concepts from other datasets, thus taking advantage of relevant data in other datasets from the Linked Open Data cloud, such as DBpedia , WorldBank[1], FLORA dataset[2] to name a few.
- Permits reasoning over the knowledge base and infer new, not explicitly stated facts and relations, support knowledge retrieval and querying using graph-matching languages (e.g. SPARQL).

Reusing already established ontologies, taxonomies and vocabularies provide means for better understanding of the concepts of the knowledge base. For this reason, it is a common practice to align semantic concepts to other recognised ontologies in order to improve interoperability. As a part of the knowledge representation within this thesis, we create a semantic model that captures extracted data, their relations and properties and align concepts to other established ontologies, such as FIBO, DBpedia, and others.

Creating financial ontology usually requires merging many different data sources into one coherent dataset. Sometimes various details are spread across multiple data sources, where each dataset contain different properties of the same concepts. In knowledge modelling, this

[1] http://worldbank.270a.info/, last accessed: 24/04/2017

[2] http://nadir.uc3m.es/flora-interface/. Last accessed: 24/04/2017

is called ontology alignment or concept mapping. For example, one company is using its own knowledge base containing companies and securities. When trying to aggregate data from another source, such as quarterly reports from EDGAR[3] system (Electronic Data Gathering Analysis and Retrieval System), one must ensure that the connections within ontology are accurately mapped to the concepts provided in the SEC fillings. Ideally, both parties could follow Linked Data principles and agree on the common URIs for all the concepts. This, however, is still a distant future. The biggest data linking hub, DBpedia, contains millions of entities and is a common reference for Linked Data concept linking, but it is still missing many entities and properties, especially in a very domain-specific area, that is Spanish stock market. Therefore there is still need for concept mapping in order to merge other relevant data. If it were possible to use a common and unique property, such as ticker symbol or unique ISIN (International Securities Identification Number), the task of concept linking would be straightforward and fully automatic. In practice, we more often face the situation where we need to employ other techniques, such as fuzzy concept label matching, ontology-based reconciliation, and even more sophisticated graph-based or machine learning techniques (Soru and Ngomo, 2014). Automatic, unsupervised methods for concept reconciliation tend to have a margin of error and in order to achieve a high level of accuracy, the whole process needs to be revised by human in order to correct any remaining mistakes or misalignments.

In order to create taxonomy of Spanish companies, we merged data coming from the following sources:

- La Bolsa de Madrid (Madrid Stock Exchange) providing up-to-date information on the list of companies traded on the Spanish stock market, their stock names, ticker symbols, stock prices, stock indices, industrial sectors and further information related to stock operations (dividends, stock splits, reverse stock splits, new equity offerings),

[3] https://www.sec.gov/edgar.shtml. Last accessed: 24/04/2017.

- CNMV providing further details on public companies, their NIFs[4], and most of all: public announcements for investors ("hechos relevantes").

Spanish stock exchange lists around 3406 companies[5], however in the context of this work we are not particularly interested in all of them. In the case of the Spanish market, we are mostly focused on relatively big companies, with enough liquidity and available for an average investor. Also, the biggest companies are ones that have relatively good press coverage.

We have also added several small and medium companies (from the "MAB Expansión" index) due to its possible future impact. We rule out most penny stocks, open-end investments companies (also called SICAVs), Real Estate Investment Trust (SOCIMI – Sociedades Anónimas Cotizadas de Inversión Inmobiliaria) and other less commonly traded financial instruments.

We have narrowed down the list of tracked companies to 209 companies from the following markets:

- Mercado Continuo: 133 companies,
- Latibex: 24 companies,
- MAB Expansión: 28 companies,
- other traded stocks (commonly referred to as "Parque"): 24 companies.

Apart from modelling the stock exchange indices' constituents, the semantic model needs also to represent financial events extracted from the unstructured sources. Together with taxonomies of related companies, relevant actors, events taxonomy, temporal and monetary units they will form the semantic model for financial events tracking. To provide interoperability, the semantic model will be aligned with established vocabu-

[4] Número de Identificación Fiscal – Taxpayer indentification number

[5] As of March 2017, according to "Revista de Bolsas y Mercados Españoles, Estudios y Publicaciones, Estadísticas", https://www.bolsasymercados.es/esp/Estudios-Publicaciones/Estadisticas. Last accessed: 21-04-2017.

laries and ontologies, especially the FIBO that is becoming business conceptual ontology standard. The extraction process will use the semantic model to create a semantic financial dataset.

## 5.2 *Relevant financial events*

Among myriads of possible events that might affect the risk and value of a company, we select the most promising to be further extracted from data. Based on domain knowledge, previous work and availability of data we choose a set of financial events that constitute our high-level features to be extracted. We aim at the most relevant and having a potentially largest impact from the point of view of risk analysis.

While typically one of the most important events is earnings announcement, there are also many other events that substantially influence the stock price and the associated risk of investment. Another example gives 67 types of different events (Antweiler and Frank, 2006). In the work of Sprenger, Sandner, et al. (2014) we find six events categories (with their corresponding subcategories), those are the following: (i) Corporate Governance, (ii) Financial Issues, (iii) Operations, (iv) Restructuring Issues, (v) Legal Issues and (vi) Technical Trading.

Tracking too many different categories is a difficult task for many reasons: (i) need to provide a corpus that covers all of the features in sufficient quantity, (ii) need to manually classify many types, (iii) classifier with too many fine-grained classes is less accurate. Other limitations of distinguishing between higher amounts of types have been described in (Antweiler and Frank, 2006; Sprenger, Sandner, et al., 2014).

Another class of events has been derived from 8-K reports that are mandatory filled by U.S. public companies on an occasion of important occurrence that might influence company results or investors' decisions (H. Lee et al., 2014). Those reports are equivalent to the Spanish *hechos relevantes* that are published in the CNMV official registry.

Based on the previous work we gather a list of event, organised into categories that are feasible to track. Those events are highly relevant and typically covered by news and many of them reported in the *hechos*

*relevantes*. Table 5.1 presents main event categories that are feasible to be tracked and important from the investor standpoint. Please note that the structure might also resemble the CNMV categories[6]. This list is, however, more tailored to the needs of this particular work. It extends the list of events that can be found in previous work, e.g. (Antweiler and Frank, 2006; Ryan and Taffler, 2004).

Many of those events are reported in an official way through the market regulator. They can be found in both CNMV reports and news sources.

Table 5.1: Classification of financial events. Equivalent Spanish terms are given in italics.

| Category | Subcategory | Event |
|---|---|---|
| **Corporate Governance**<br>*Gobierno Corporativo* | **Board & shareholders**<br>*Consejo de administración y convocatorias oficiales* | **Board composition changes**<br>*Composición del consejo de administración*<br>**Shareholders' meeting announced**<br>*Convocatorias y acuerdos de Juntas y Asambleas generales* |
| | **Company status**<br>*Modificaciones estatutarias* | **Company status change**<br>*Modificaciones estatuarias*<br>**Shareholders' rights changes**<br>*Cambios de control*<br>**Yearly corporate governance reports**<br>*Informes anuales de gobierno corporativo* |

[6] https://www.cnmv.es/portal/HR/BusquedaHR.aspx

| Category | Subcategory | Event |
|---|---|---|
| | **Key Personnel**<br>*Nombramientos* | **New CEO is appointed**<br>*Nombramiento de nuevo director ejecutivo*<br>**Key management change**<br>*Composición de órganos de gestión y control*<br>**CEO resigns**<br>*Dimisión de director ejecutivo* |
| **Financial Instruments**<br>*Instrumentos financieros* | **Stock**<br>*Acciones* | **Dividend Announcement**<br>*Anuncio de dividendos*<br>**IPO**<br>*Oferta pública de venta de acciones (OPV)*<br>**New Stocks Issue**<br>*Oferta pública de suscripción de acciones (OPS)*<br>**Stock Split**<br>*Desdoblamiento de acciones*<br>**Reverse Stock Split**<br>*Agrupamiento de acciones*<br>**Stock Buyback**<br>*Programa de recompra de acciones* |
| **Financial Situation**<br>*Situación financiera* | **Periodic Reports**<br>*Información financiera* | **Earning reports**<br>*Información sobre resultados*<br>**Sales and performance reports**<br>*Otros informes de ventas y rendimiento*<br>**Company debt management**<br>*Préstamos, créditos y avales* |
| | **Ratings**<br>*Calificaciones crediticias* | **Company rating changes**<br>*Revisión de calificaciones crediticias* |

| Category | Subcategory | Event |
|---|---|---|
| | **Analysts' Forecasts**<br>*Opinión de analistas* | **Analysts earnings estimates**<br>*Estimaciones de beneficios* |
| | **Other**<br>*Otros* | **Bankruptcy**<br>*Quiebra* |
| **Human Resources**<br>*Recursos humanos* | | **Hirings**<br>*Contrataciones*<br>**Layoffs**<br>*Despidos*<br>**Substantial employment changes**<br>*Cambios substanciales en la plantilla* |
| **Legal Situation**<br>*Situación legal* | **Legal Issues**<br>*Problemas jurídicos* | **Company under investigation**<br>*Empresa bajo investigación*<br>**Class action**<br>*Proceso de conflicto colectivo*<br>**Company Sanctioned**<br>*Empresa sancionada*<br>**Fraud**<br>*Fraude*<br>**Trading suspension**<br>*Suspensión de cotización* |
| | **Law issues**<br>*Cuestiones de derecho* | **Fiscal law changes**<br>*Cambios de legislación fiscal* |
| **Market Environment**<br>*Entorno del mercado* | **Competition**<br>*Competidores* | **New competitor**<br>*Entrada de nuevo competidor*<br>**Competition change**<br>*Cambio de competidores* |
| | **Raw Materials**<br>*Materias primas* | **Raw material price change**<br>*Cambio de precio de las materias primas* |

| Category | Subcategory | Event |
|---|---|---|
| | **Other**<br>*Otros* | **Major disaster**<br>*Grandes desastres* |
| **Products and Services**<br>*Productos y servicios* | | **New product release**<br>*Lanzamiento de nuevo producto*<br>**Patent issued**<br>*Concesión de patente*<br>**Issues with product**<br>*Problemas con producto*<br>**Product market share changes**<br>*Cambio de cuota de mercado* |
| **Strategic Operations**<br>*Operaciones estratégicas* | **Transformation**<br>*Transformaciones* | **Acquisition**<br>*Ofertas públicas de adquisición (OPA)*<br>**Company restructure**<br>*Reestructuración de empresa*<br>**Liquidation and dissolution**<br>*Liquidaciónes y Disoluciones*<br>**Merger**<br>*Fusiones*<br>**Spin-off**<br>*Escisión* |
| | **Agreements**<br>*Acuerdos con terceros* | **Strategic alliance**<br>*Acuerdo estratégico*<br>**Collaboration**<br>*Colaboración*<br>**Joint-venture**<br>*Joint-venture* |

| Category | Subcategory | Event |
|---|---|---|
| | **Expansion**<br>*Expansión* | **Business expansion**<br>*Expansión de negocio*<br>**Business scope changes**<br>*Cambios de áreas de negocio* |

## *5.3 Common semantic event model*

Capturing and representing financial events implies the existence a common semantic model, that could resemble as good as possible all aspects and properties of such act. Such model can be either created or reused, based on individual requirements and availability and suitability of other existing models.

In modelling real-world events we would like to be as genuine as possible, but also provide a simple, yet sufficient way for expressing facts. For example, in journalism, every story should answer the basic questions that can be shortened into five Ws: What, When, Where, Who and hoW. This is the way of simplifying the most important aspects of a reported story and ensure the minimum baseline for reported information. Based on that we look at a minimum set of information enough for representing a financial event.

We start by defining a conceptual framework for event representation for financial events extraction. Having in mind most common attributes and their importance, we define an event as a 6-tuple of a form:

$$e_i, t_k, \ p_j \ , \ r_j \ , l_m, s_n$$

where $e_i$ is an event type (**what happened**), $t_k$ is **when** (defined as time or interval) such event occurred, $p_j$ is a multiset of parties participating in the event (**who is involved**), $r_j$ is list of roles that correspond to their participants (**what role she/they play**), $l$ is a location where the event took place (**where**), and $s$ is a source of information about the

event (**how we know**).

While this definition provides a good general overview of the event characteristics, it is still lacking details on how to unambiguously map a real-world event into a specific representation. Such representation that could not only provide abstract concepts that fit definition but also allow to properly and faithfully model such events. For this, we will use ontology modelling in order to precisely define the knowledge representation for our extracted events.

We studied previous work on this topic in order to align with existing efforts, and following the good practice of ontology reuse. The event modelling with ontologies is well-researched and broad topic, covering many domains with different applications, such as generic news and situation events (Segers et al., 2015), news storylines (Wilton et al., 2013), business news (Lösch and Nikitina, 2009), corporate products and events (Kakkonen and Mufti, 2011), incidents events (Fani and Bagheri, 2015).

From the point of view of available ontological resources, the situation is different, as only a few ontologies are freely available. Not many authors publish or otherwise share their ontologies or they have already disappeared from the web by the time of writing of this thesis. Another issue is that the fundamental property of ontology is a shared conceptualisation, agreed among domain experts. This implies a certain consensus on what is the preferred way to represent data, an acceptance of some canonical form, and adoption of a most established model.

The most notable available and existing models, that are still relevant and suitable for event modelling are:

- The OpenCyC ontology, containing a generic classes for modelling "Events and Situations" (Matuszek et al., 2006)

- DBpedia event model, with the `event` class[7] and EAV[8] model (see Section 2.2). This model and the DBpedia Live dataset is of a growing importance due to near real-time processing of edits and facts extrac-

[7] `http://dbpedia.org/ontology/Event`

[8] EAV means Entity Attribute Value

tion from the original Wikipedia. (Magnus Knuth Jens Lehmann and Sack, 2015)

- Schema.org provides a generic event class for expressing things that happen at certain time and place, such as concerts, exhibitions, etc. `https://schema.org/Event`
- DUL[9], the DOLCE+DnS Ultralite ontology is a evolution of the original DOLCE upper ontology (Gangemi, Guarino, et al., 2002), simplifying terms and including Descriptions and Situations (DnS) ontology (Gangemi and Mika, 2003).
- OpenCalais proprietary ontology for automatic news annotation[10]
- PROTON (PROTo ONtology) lightweight upper ontology for basic text annotation and Information Extraction tasks[11]
- The FIBO Foundation ontology that provides general definitions contains classes such as `Occurence` or `PartyInRole` useful for defining event properties. The ontology is still in development and more coverage of different financial domains is still to be seen.

Based on the study of existing ontologies and vocabularies, and given our requirements we provide a simplified core EAV model for event definition, aligned with Schema.org and DUL models, and mapped to FIBO and DBpedia. In this sense, we are not aiming at creating a new ontology, but rather extend existing models, in order to provide necessary means for modelling our dataset.

Figure 5.1 shows a conceptual model for representing extracted events. Concepts represented in green and blue are part of our financial semantic dataset, created in the process of information extraction. Both classes and relations have direct mappings to equivalent concepts as defined by

[9] `http://ontologydesignpatterns.org/wiki/Ontology:DOLCE+DnS_Ultralite`. Last accessed: 11/04/2017.

[10] `http://www.opencalais.com/wp-content/uploads/2015/06/Thomson-Reuters-Open-Calais-API-User-Guide-v3.pdf`. Last accessed: 11/04/2017

[11] `http://ontotext.com/products/proton/`. Last accessed: 11/04/2017

previously mentioned well-adopted ontologies. We also provide mappings to several DBpedia entities and properties, following the Linked Data approach.

Table 5.2 shows the corresponding classes of our dataset and other ontologies and vocabularies. Our extension provides three taxonomies (shown in blue): Event taxonomy, Participants (with Spanish companies ontology) and Roles. All three taxonomies are also aligned with the existing FIBO specifications, in this case, the FIBO Foundations ontology.

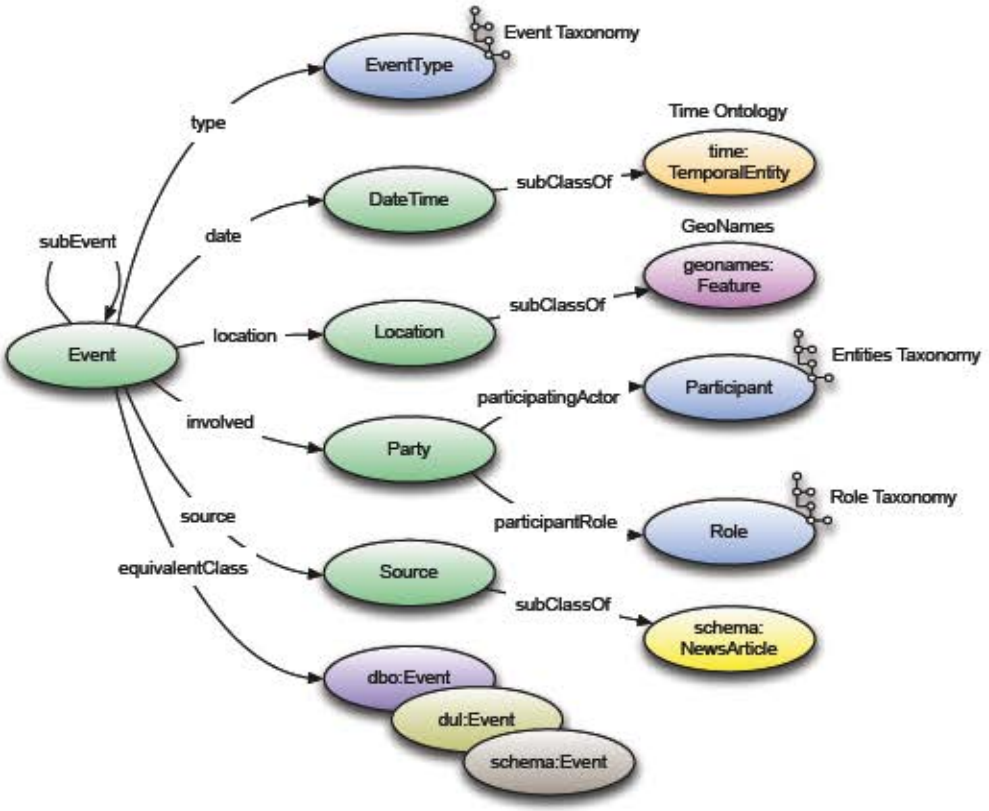



Figure 5.1: Extending the event model.

We define vocabulary classes through by their relationship with other ontologies:

- The `Event` class is central part representing the actual event, as extracted from the news. It maps to DBpedia `dbo:Event`, Schema.org `schema:Event` and `DUL:Event`. In case of a complex event, where an event is composed of more sub-events, we establish circular relation

`subEvent` to relate a super event with its sub-events. This property subclasses `schema:subEvent`, and its inverse property: superEvent subclasses `schema:superEvent`.

- Within the event model we are using Time Ontology (Cox and Little, 2012) to define temporal aspect of the event. It can be either an concrete date (i.e. `time:Instant`) or a period (`time:DateTimeInterval`).
- The `Location` maps to a GeoNames top level concept that can be further specialized into concrete places (such as cities, countries, etc.).
- The `Source` class describes a source news where this event has been extracted from. This class subclasses `NewsArticle` from Schema.org, and individuals can use also Dublin Core metadata (Dublin Core Metadata Initiative, 2012) to provide additional fields, such as author, date, publisher or URL.
- The `Party` class establishes a relation between an event, a participating party (or parties) and their roles. As each entity (be it a company or a person) might have a different role in different events, an intermediary class is required to accurately model such situations. While Schema.org, DUL or DBpedia ontologies all provide some way of representing participants. They, however, follow more flat structure (like EAV), where the distinction of roles is not directly possible without extensive use of property hierarchies. Other ontologies (e.g. FIBO) recognize this pattern and provides a more expressive model for such representation.

Although Schema.org or DBpedia vocabulary provide less expressiveness, we can still at the same time provide a compatible, "simplified" view, where a role is simply not present. This is due to the fact that the dbo:participant can be inferred from the chain of `e:involved` and `e:participatingActor` properties through the Object Property Chains (W3C, 2012). The same applies to the DUL ontology and the following relation: `dul:isParticipantIn`.

The classes involved in the event model can be further extended i.e. through other, more specialised ontologies. The model provides means

for that by specifying classes where a more specific class can be sub-classed. A bottom-line list of most basic imports is shown in Table 5.3.

| Dataset concept | Foreign Mapping |
|---|---|
| e:Event | dul:Event |
| e:Event | dbo:Event |
| e:Event | schema:Event |
| e:DateTime | time:TemporalEntity |
| e:ComplexEvent | schema:EventSeries |
| e:EventType | dul:EventType |
| e:subEvent | schema:subEvent |
| e:superEvent | schema:superEvent |
| e:Source | schema:NewsArticle |
| e:involved/e:participatingActor | dbo:participant |
| e:involved/e:participatingActor | e:eventParticipant |

Table 5.2: Event model mappings. The namespaces are: `dbo` – DBpedia Ontology, `dul` – DUL upper ontology, `schema` – Schema.org. Our dataset namespace is `e:`

| Ontology | Role |
|---|---|
| FIBO | dul:Event |
| Geonames | dul:Event |
| SKOS | dul:Event |
| DBpedia | provision of mappings |
| DBpedia | provision of mappings |

Table 5.3: Ontology imports

## 5.4 *Taxonomies*

The taxonomies that extend the event model are used to express main characteristics of extracted financial events. The following subsections give an overview of each taxonomy and their role in modelling extracted events. It provides a more general description in order to familiarise the

reader with the big picture and explain the most important concepts of the semantic model. More fine-grained details can be found in Appendix C.

### 5.4.1 *Role taxonomy*

The notion of company role in the event can be defined in 2 ways: (i) either by defining a relation between classes `e:participantRole` and `e:participantActor` to the `e:Party` or (ii) by directly relating the `e:Event` with `Participant` through the taxonomy of predicates that describe the participant's role. Both representations are equivalent in terms of transmitted knowledge, as one representation can be inferred from another e.g. through the property chain reasoning rule of OWL2. However, both representations have some fine-grained differences. Figure 5.2 presents both approaches to relation modelling. The "Participant-in-Role" needs an extensive use of additional nodes that relate an event with a participant. This allows attaching additional event-specific properties to participants' roles in the event. On the other hand, the "Role-as-Relation" express the same information using a hierarchy of properties that relate an event with a party. This is closer to the EAV approach that is extensively followed in many LOD datasets. In further mentions, we will favour the EAV-like approach.

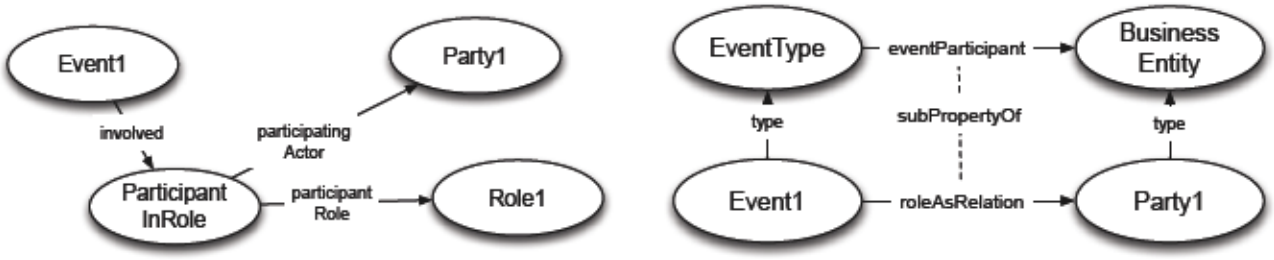



Figure 5.2: Two approaches to role modelling. On the left the "Participant-in-Role" variant. On the right the "Role-as-Relation" variant following the EAV approach.

In the case of the former, the matching between some "Event1" and all involved Parties and their roles can be then discovered through the

conjunction of the following relations:

$$\begin{aligned} &\text{involved}\ \text{Event1}, ?\text{Party} \\ &\text{participatingActor}\ ?\text{Party}, ?\text{Participant} \\ &\text{participantRole}\ ?\text{Party}, ?\text{Role} \end{aligned}$$

This can be obtained from our dataset through the following SPARQL query:

```
PREFIX alfredo: <http://nadir.uc3m.es/alfredo#> .
PREFIX data: <http://nadir.uc3m.es/alfredo/dataset#> .
PREFIX rdfs: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

SELECT * WHERE {
        data:Event1 alfredo:involved ?party .
        ?party alfredo:participatingActor ?participant .
        ?party alfredo:participantRole ?role .
}
```

Listing 5.1: SPARQL query

For the direct relation, the resulting graph resembles more the EAV structure, such as DBpedia. The navigation is easier through the omission of the `Party` node. The semantics of the event, as such is however maintained, through the use of taxonomy predicates and `EventType` taxonomy.

Finding participants involved in the event is therefore looking for all triples that satisfy the following relations:

$$\begin{aligned} &?\text{relation}\ \text{Event1}, ?\text{Participant} \\ &\text{subPropertyOf}\ ?\text{relation}, \text{eventParticipant} \end{aligned}$$

In SPARQL this query is expressed as follows:

```
PREFIX alfredo: <http://nadir.uc3m.es/alfredo#> .
PREFIX data: <http://nadir.uc3m.es/alfredo/dataset#> .
PREFIX rdfs: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

SELECT * WHERE {
        data:Event1 alfredo:relation ?participant .
        ?relation rdfs:subPropertyOf alfredo:eventParticipant .
}
```

Listing 5.2: SPARQL query for relation retrieval

As shown by examples above, the Role taxonomy is, therefore, a taxonomy of classes for the longer version or a taxonomy of properties for the shorter one. In the dataset, we use the shorter, EAV-like version, that makes more use of RDFS reasoning and is more concise. The resulting SPARQL queries are also faster.

### 5.4.2 *Event type taxonomy*

The taxonomy of financial events has been modelled after the classification given in Section 5.2. The taxonomy is a hierarchy of classes describing company events starting with the root class `EventType`, thus extending the event model with a hierarchy of concrete financial events. Those classes are to indicate a concrete type for an `Event`. Figure 5.3 gives an overview of the subclass relations.

Further description of each class and sub-class hierarchy is presented in Appendix C.1.

## 5.5 *Event representation example*

In order to give a computable and highly interoperable form for the high-level information extracted from the news, we employ all the previously mentioned elements as building blocks of what we call a semantic event representation. The semantic model explained in the previous sections aims at representing all extracted facts in order to construct

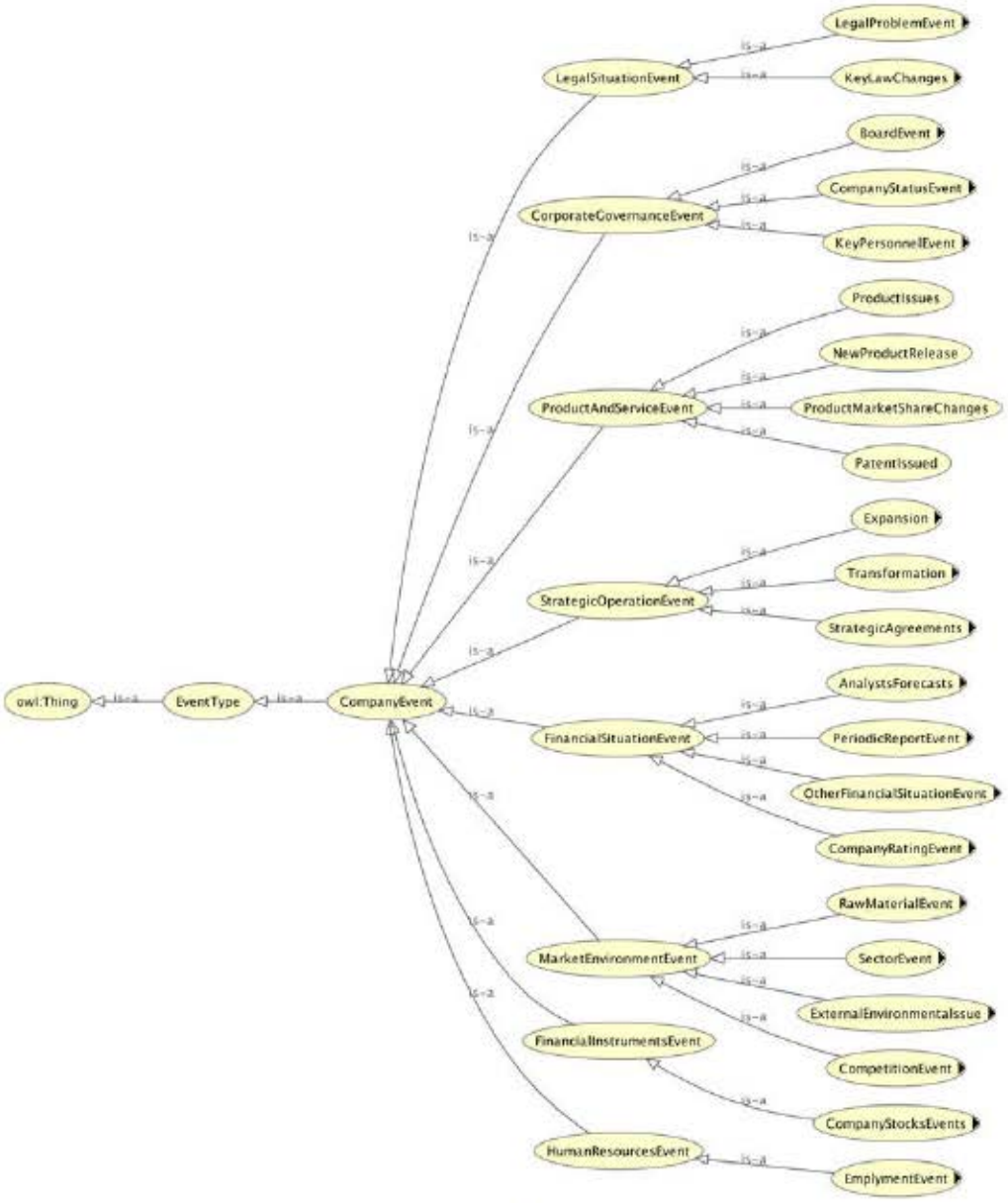



Figure 5.3: Overview of the taxonomy of event types. For brevity, not all subclasses has been shown.

the semantic knowledge base that will further be used in process of financial decision-making. Figure 5.4 shows semantic representation of a single event. In this case this an announcement of company acquisition (Spanish: *OPA*). Each colour represents a different fragment of the ontology. Rounded nodes depict concepts and the rectangular node is a literal node. As seen in this example, we also follow the EAV-like approach to the role modelling (as in "Role-as-Relation" example from Section 5.4.1.) `FerrovialBroadspectrumOpa` is the central node that represents this event. It is of a class `Acquisition` from the `EventType` taxonomy (see Section 5.4.2). A simplified fragment of this taxonomy is visible (in purple) in order to show that it extends e.g. `schema:Event` type (Schema.org types are in brown). The `FerrovialBroadspectrumOpa` node also contains some metadata that was extracted from the source document. Both, the central node and the `SourceArtcile1` are representing information directly extracted from news (in green colour). The metadata attached are (i) date and (ii) source article. Date is expressed through the `time:TemporalEntity` of the OWL Time ontology (Cox and Little, 2012) (in asparagus colour). The most important are relations between `FerrovialBroadspectrumOpa` and `Ferrovial` and `Broadspectrum` depicted as the following arrows: `acquiringParty` and `acquiredParty`. Both relations, characteristic for the `Acquisition` type point to: (i) an entity that is the acquiring party and (ii) acquired party accordingly. `Ferrovial` and `Broadspectrum` are companies participating in the event, as described by relation. Both are part of the ontology that is used for ontology-based NER, so that they can be detected in the text. Both are of type `Company` that maps to the FIBO Foundations ontology, and the Corporations taxonomy in particular. A fragment of the FIBO taxonomy has been presented for the reader's convenience (in light blue). It is also shown that both relations `acquiringParty` and `acquiredParty` are specialisations of the generic `eventParticipant` property, following the Role taxonomy model (see Section 5.4.1 and Appendix C.2). Other financial events are represented in a similar way.

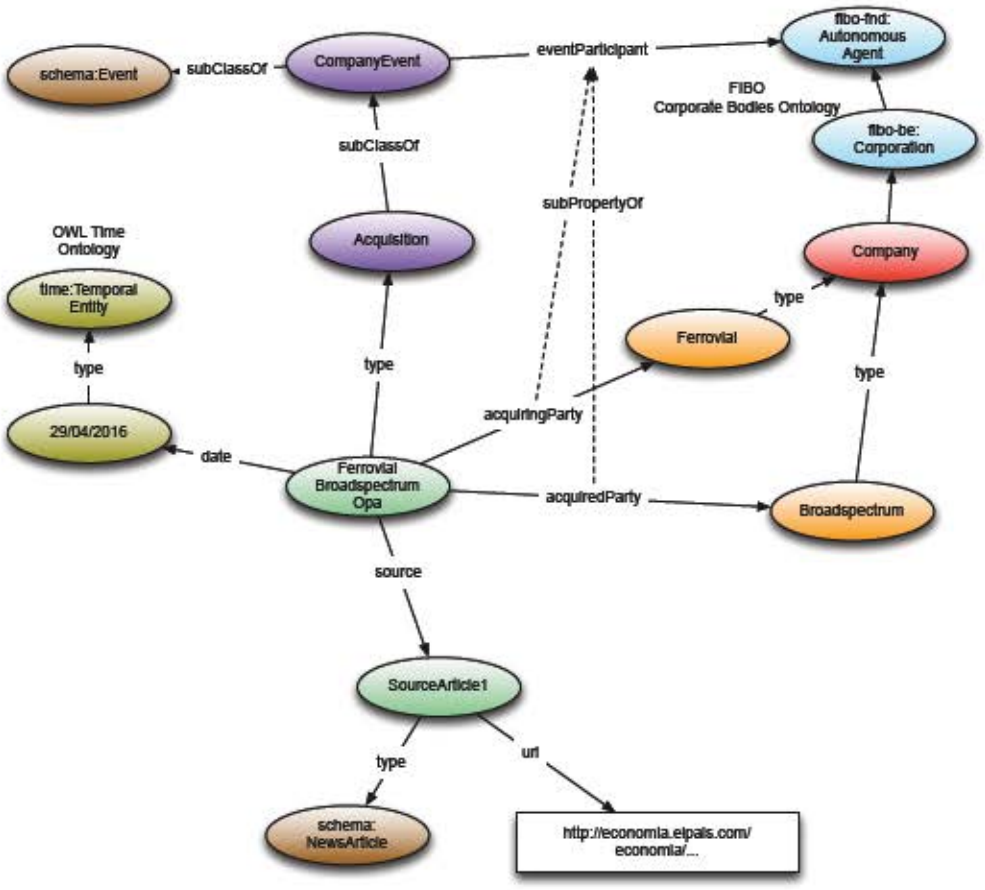



Figure 5.4: Example of semantic event representation.

## 5.6 *Note on the modelling approach*

The semantic modelling work described in previous chapters is a result of an iterative approach to ontology engineering. Instead of following heavyweight methodologies that were not suitable for this task, we rather loosely followed a set of paths that were the best fit for the experiment-driven, iterative process. The knowledge engineering was largely inspired by the NeON methodology (Suárez-Figueroa et al., 2012) due to its flexibility. Rather than favouring one canonical approach, it offers several scenarios for ontology construction and defines various different paths that can be followed, depending on the actual requirements. This is in line with our observation, that there is no single approach that could match every possible scenario (Pfleeger, 2009). In the adopted

modelling approach we were mostly focused on:

- Reusing, Merging, and Re-engineering Ontological Resources — in order to follow the best practices in ontology reuse, avoid reinventing the wheel and make use of already established vocabularies. The limitations of the resources available required us to provide domain-specific extensions (more in Section 5.3). We also provided alignments to other common semantic vocabularies that open the possibility for future merging of semantic datasets.

- Reusing and Re-engineering Non-Ontological Resources — At some point in time a sole class hierarchy is not enough. Especially when, in the course of experiments, we require some concrete instances. For example, in the case of ontology-based NER, we need instances together with their names, labels, etc. This is the scenario that was followed in order to start populating ontology with some instances in order to carry on the experimentation process.

- Straightforward ontology specification — In other cases, when it is not possible to reuse any ontological artefact, we need to create our own ontological resource. Our taxonomies that extend the event model were created following this scenario. We started with knowledge acquisition and performed few iteration between conceptualisation and implementation, where we used current experiment results as a feedback for next iteration.

Creating a complete domain ontology is a long process. The bigger the ontology, the more complex it becomes to conceptualise and formalise all the requirements, and more difficult the actual implementation. Instead of trying to conceptualise the whole domain, we rather followed a more lightweight approach. We designed a set of smaller semantic models, highly specialised in a concrete task. They comprise event model, accompanying semantic taxonomies and alignments with other semantic resources. This way we could iteratively extend taxonomy to fit the experimentation process and improve the overall model at the same time.

## 5.7 Conclusions

The "Semantic Modelling" is the second step in the roadmap toward the automated decision-making process. Previously we identified and prepared source news data and we stopped. This is because any further Information Extraction task can not be made without identifying the actual information we aim to extract. Therefore we performed an analysis of the *features* of interest for the financial decision-making process. This step is also closely tied to the target representation of our extracted facts. And this is precisely when the semantic modelling starts.

We presented our main motives for employing Semantic Web technologies in the knowledge modelling and defined its expected role in the overall process. As the semantic modelling is inseparably connected to the domain where it is used, we contextualised the decision-making process in the financial domain. After that, the most relevant financial events were identified. Those events are the "core" around which all the modelling happens. The semantic model consists of a common financial event model and its accompanying taxonomies that provide more specialised concepts for expressing different real-world situations.

We are fully aware that nowadays no ontology is created entirely from scratch. We abide by the best practices for ontology engineering, such as following the LOD approach, reuse of established vocabularies, provide mappings to other datasets. In this process, we were mostly inspired by the NeON methodology (Suárez-Figueroa et al., 2012).

As a result, we have prepared a model for expressing knowledge about financial events. Now the DSS consists of two unconnected pieces: a huge news corpus from the "Data Acquisition" step and an idea on how to model events from the "Semantic Modelling" step. But there is still a missing connection between both. The explanation on how to bridge this gap is given in the next chapter.

# 6 *Financial events extraction*

This chapter describes the natural language processing and machine learning approach to financial event extraction and classification. Based on the previous steps we aim at processing the corpus text created in Chapter 4 with the purpose of extracting *high-level features* described in Chapter 5. This covers the whole process that comprises corpus creation, annotation and classification. By employing novel classifications methods using deep neural networks together with semantic taxonomies we improve classification of events and relations. In this sense, this chapter bridges the gap between unstructured data and ontologies: the extracted facts from the unstructured data are further modelled using the semantic model devised in the previous chapter. All facts extracted from our text corpus are later stored in the semantic knowledge base and following the model devised in Section 5.4. The resulting knowledge will be further used in the process of decision-making. This chapter shows how we approach the information extraction in order to extract semantic facts.

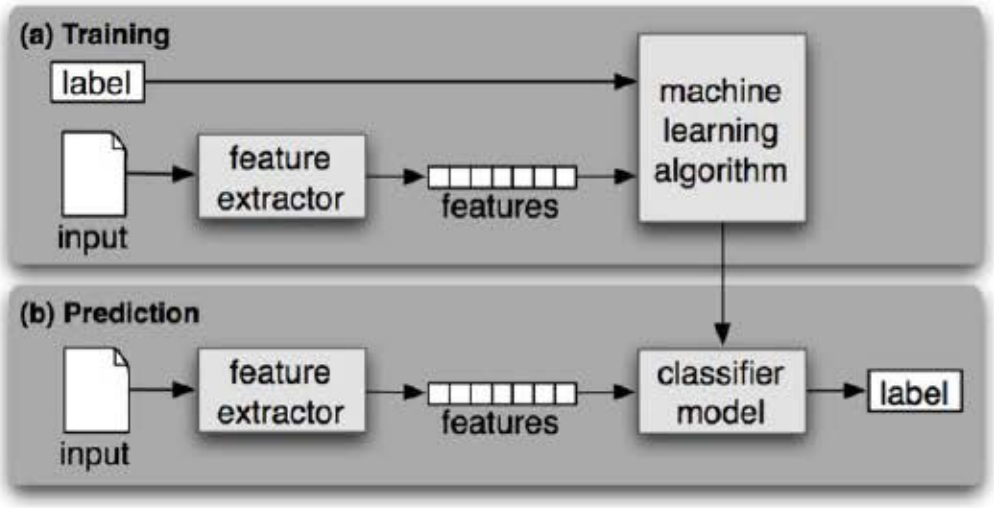



Figure 6.1: Supervised training and classification overview (Bird et al., 2009).

## 6.1 *Feature analysis and corpus annotation*

Classification of financial events is performed in the framework of machine learning. In the scope of this work either supervised, semi-supervised and unsupervised methods are considered. The use of machine learning techniques requires proper definition and preparation of the training set. In the case of supervised learning, a typical scenario requires using a gold standard, or (if such corpus does not exist in the domain) to manually annotate data corpus in order to provide training set. The annotated (labelled) data is used to train machine learning algorithm and create a classification model that is further used to classify new data automatically (see Figure 6.1).

In the domain of natural language processing, providing a labelled corpus for supervised learning is performed in manual a process called annotation. It consists of attaching labels to appropriate chunks of texts (words, phrases, sentences etc.) in order to indicate selected relevant features that should be captured by the machine learning algorithm.

An example of a relevant event is a change of CEO of some company. Annotating a NewCEO event has been presented in Figure 6.2. In this case, we annotate a binary relation between a Person and an Organisation, which states that this person became a new CEO of mentioned


Figure 6.2: Manual sentence annotation for relation extraction task. The relation between named entities, in this case PERS and ORG is established manually by the annotator.

company. We can define this relation as a binary relation as follows:

$$NewCEO(Organisation, Person)$$

Annotation of features for relation extraction as presented above is crucial, however not sufficient for successful information extraction task. In the traditional machine learning approach, the annotation PERSON and ORGANISATION is one of many features that are desired in order to provide decent accuracy for classification. Each feature is a piece of information that can help the algorithm to learn complex patterns and properly classify as many testing samples as possible. Those features are typically other linguistic aspects of the given text, such as part-of-speech (POS) labels, entity mentions (through the NER: Named Entity Recognition process) word dependencies (and dependency trees) distance between key words and many others. In the previous example (Figure 6.2) apart from PERSON and ORGANISATION, we can also see Spanish POS tags above each word. All such features are used to train a model for relation extraction. In the context of relation extraction, a detailed description of features that yield good results is presented in (Björne et al., 2011).

Entity mentions play a special role among annotations, due to the fact that they directly indicate concepts involved and are constituents of many relations. Their sole presence might be a good indicator of candidate phrases for relation extraction task. Figure 6.3 shows an example of NER tagged sentence together with co-references (i.e. "presidente",

"Dimas Gimeno Álvarez", "quien" and "su" all refer to the same person). There are various different approached to NER (Nadeau, 2007) based on techniques used. In the case of Business Intelligence we face with two problems: on one hand, we prefer a controlled list of entities (e.g. company names list which is finite and usually quite manageable), in order to be sure that we never miss a single mention. On the other hand, we are interested in entities that we do not know upfront, such as people given names, product names dates, etc. Both cases are quite different and in most cases they can not be solved with the same approach. In the first case, the most suitable is extraction based on some predefined list, such as a *gazetteer*. Then the whole process boils down to pattern matching in the text. We can also use some more sophisticated inputs, such as ontologies or semantic vocabularies to perform an ontology-based NER in a similar manner. The more generic approach to NER, when it is not feasible to provide a closed-list of entities, are based on machine learning models. While those model-based NER approaches can successfully classify a wide range of previously unseen entities, the error rate is also higher. The further description of the information extraction process through NLP techniques for ontology-based NER and model-based NER is given in section 7.1. Note that when ontology-based NER is performed, an annotated entity already has an existing concept in the ontology. When model-based NER is applied, the new annotation points to a new concept that is unlikely to exist in the ontology. Such concepts may be correct or not, thus a human intervention is necessary. This thesis does not cover various aspects of ontology evolution, as it is out of the scope of this work. In this respect, we rely on manual concept curation when necessary.

Many highly specialised tools support the task of corpus annotation and provide user interfaces to accomplish word and phrases labelling, relations, chunking, dependency annotations and other. After evaluation of existing annotation software, the choice of BRAT Rapid Annotation Tool[1] has been made, due to the presence of a web-based GUI, possibility

[1] BRAT project URL: http://brat.nlplab.org/index.html. Last accessed: 2016-10-11.

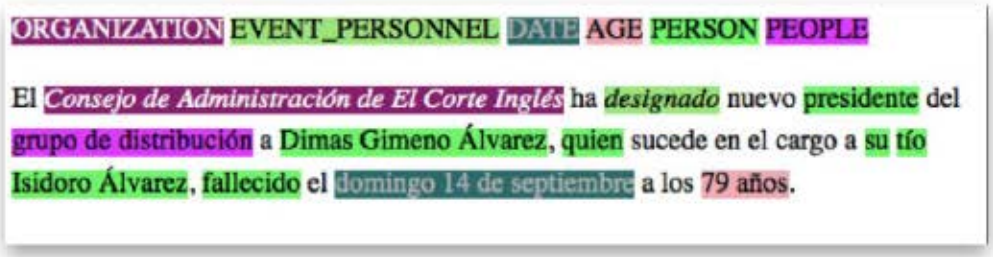



Figure 6.3: Sentence with named entities (together with their co-references) annotated.

for annotating n-ary relations (events), and possible interoperability with Apache UIMA framework (Stenetorp et al., 2012).

## 6.2 *The bootstrapping technique*

Text classification tasks very often require enough information for training statistical algorithm with the desired accuracy. This implies the use labelled corpora as a ground truth for training and evaluation tasks. On many occasions, the linguistic resources are scarce and or not available at all. This is when it is necessary to manually create a new corpus. Creating corpora for Natural Language Processing is a very tedious and time-consuming task, involving manual annotations of large amounts of texts. Text annotations can be sometimes very detailed, including labelling of many features, leading to even more complexity of the whole task. For instance, the classification of the NewCEO relation requires preparation of positive and negative examples. That is a set of sentences where such relation occurs (positive) and a set where it does not (negative). The two-class classification is called binary classification problem.

In many situations, we can overcome the problem of corpus creation. Bootstrapping is a technique that helps to expand corpus from a few examples (called *seeds*). The initial seeds are used to find more suitable ones. The general idea is based on Dual Iterative Pattern Relation Extraction (DIPRE) algorithm (Brin, 1999) further extended into the Snowball algorithm by Agichtein and Gravano (2000). The central idea can be summarised as:

1. Start with an initial list of carefully chosen examples (seeds).
2. Use those examples to gather more examples alike.
3. Evaluate chosen items and select the ones that are best matching the pattern.
4. Go to step 2 and repeat the process until enough examples have been gathered.

The process described above works very well for relation extraction tasks, where finding similar examples can be defined in a relatively straightforward way. For instance, the NewCEO relation can be bootstrapped with items presented in Table 6.1. Based on those items we can

| ORGANISATION | PERSON |
|---|---|
| Barclays | James Stanley |
| CaixaBank | Jordi Gual |
| Campofrío | Fernando Valdés |
| Euskaltel | Francisco Arteche |
| Calidad Pascual | José Luis Saiz |
| Petrobras | Pedro Parente |

Table 6.1: Bootstrap seed examples.

look for all occurrences where both entities appear very close to each other (or simply in the same sentence). We can use popular search engines to find more items by for instance using the `NEAR` keyword[2]. Based on acquired examples we further extract patterns, such as in the original Snowball algorithm, where each occurrence of two named entities is assigned a 5-tuple containing the surrounding context. The words in the 5-tuple are evaluated based on their importance (measured in terms of frequency) and grouped by a similarity function. Tuples that share common terms have higher similarity measure. This is used to induce patterns for finding further examples: the pattern obtained from this

[2] This keyword in different forms is available on the popular search sites, such as Bing or Google.

step is a centroid vector of tuples in a group Agichtein and Gravano (2000).

Other techniques were also studied in a context of relation classification, e.g. in KnowItAll (Etzioni et al., 2005) and in TextRunner (Banko et al., 2007). While they provide certain advantages to relation classification techniques, such as automatic entity extraction (KnowItAll) or discovering relations automatically (TextRunner) they also go way beyond the scope of this task.

In our case, we simplified this approach by establishing the most common pattern based on the results of the seed item list. For instance, for the *NewCEO* relation, the pattern consisted of two entity mentions: PERSON and ORGANISATION and a list of frequently occurring words, expanded with their synonyms.

## 6.3 *Information extraction pipeline for the knowledge base population*

The process of information extraction can be described as a natural language processing pipeline that is capable of analysing documents from the web and continuously extract relevant features for the financial decision support. Each stage of the pipeline is specialised to provide distinct incremental functionality that operates towards the final goal. It starts with the acquisition of source documents from the web and aims at extracting relevant information beginning with lower-level features (clean-text extraction, NLP tagging). Those lower-lever features are later used for higher-level classification tasks.

The pipeline stages for information extraction task are the following:

- Data acquisition and pre-processing components (as described in details in Chapter 4.2) with the main goal of tracking and crawling news sources and transforming them into clean text for next stages.

- Natural language processing components:

  – Segmenting – splitting text into sentences and words (tokens)

  - Lemmatising – reduce inflectional forms of words, and finding a base form for a word, in order to be able to analyse different forms of the same word as a single item.
  - POS Tagging – assigning (tagging) a part-of-speech to a word (e.g. noun, verb, adjective, etc.). This task is important for the subsequent analysis, e.g. for finding named entities (that are typically nouns).
  - Dependency parsing – analysing grammatical relation of words in a sentence and create a sentence structure representation in a form of a dependency (or constituency) tree.

- Named Entity Recognition
  - Ontology-based NER – detecting named entities by means of controlled vocabulary. The basis for this is the company list from Section 5.1.
  - Statistical, model-based NER – detecting new entities, not present in the ontology, as candidates to be included in the ontology. In the case of our linguistic pipeline we use the available Spanish models for NER (more in Section 7.1).

- Relation extraction and document classification
  - Classification models for document classification – classifying whole documents to a concrete category as a preliminary step can filter out documents that are possibly not relevant for the financial decision-making process.
  - Relation extraction – a model-based classification for relation extraction from documents and sentences. Each model is trained for a concrete relation based on prepared corpus.

Processing a stream of documents results in a series of high-level features extracted from the text, such as detected entities, related financial events, business relations, etc. Those high-level features are later semantically represented and stored in a knowledge base (triplestore). An overview of this process is presented in Figure 6.4. It depicts the main

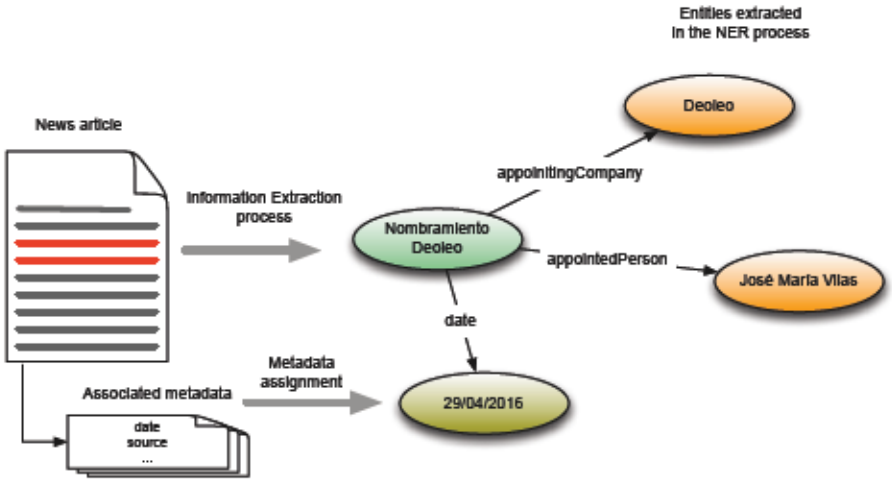



Figure 6.4: Schematic overview of the extraction process. It depicts the main goal of the information extraction pipeline for the knowledge base population.

idea of Information Extraction in order to produce semantic representation out of extracted high-level features. At this point, we take advantage of previously extracted lower level features (such as metadata) in order to lift this information to semantic level. This information forms an input for a subsequent decision support model. In the next sections, we demonstrate how we approach the Information Extraction task.

## 6.4 *Word representation in semantic vector space*

While sentence annotation and feature engineering are crucial activities of classical NLP classification techniques, the latter sometimes can be significantly reduced. The recent development in the field of artificial neural networks provides a good indication that features can be also learned in the process of machine learning when using deep neural network architectures. In this case, multi-layer neural networks can perform *feature learning* (also called *representation learning*) at their hidden layers and use them for classification in the subsequent layers. Moreover, the process of feature learning can be structured by providing more stacked layers in order to learn more abstract features. The overview of such process has been presented in Figure 6.5. The main advantage of this approach is that the hand-crafted feature engineering can be effectively

learned by the same machine learning algorithm. There are also other techniques for efficient representation learning that will be shown in the Section 6.4.

The classical information retrieval and text mining systems also have one primarily disadvantage. The input words are in principle equal to each other: a priori there is no meaning behind the symbols. Classic machine learning treats words as numeric *tokens* by assigning each word a unique number (i.e. index in a vocabulary). Then the process of text modelling operates on such defined set of tokens (numerical indexes). For instance document similarity, measured as a cosine similarity between document vectors, or ranking terms within documents using term frequency–inverse document frequency metric, are all based on this preliminary step of indexing words, and further operating on such numerical indices.

In many applications, this approach is sufficient and provide good results. However, it has many limitations. First of all, in many languages, words are often subject to inflection that result in vocabulary that contains multiple versions (inflections) of the same word. For example, English words "goes" and "going" both refer to the same verb "go", but in the process of tokenisation, they will have a different numerical index in the vocabulary. This disadvantage can be overcome by lemmatisation or stemming. Lemmatising a word means to find its *lemma* – a canonical form that will be used for all its *lexemes*. The idea behind stemming is similar, but instead of mapping all lexemes to its lemmas, we find the part of the word that remains the unchanged for all its inflected forms. For instance for words "practicing" and "practical" the stemmed form would be "practic-".

While word stemming and lemmatisation substantially reduces vocabulary size and map words to their various grammatical forms, it still does not resolve the intrinsic problem of classical word tokenisation. For instance, the following two Spanish words *jefe* and *director* have a very close semantic meaning in the context of corporate governance. Even after applying stemming or lemmatisation, we will get two different

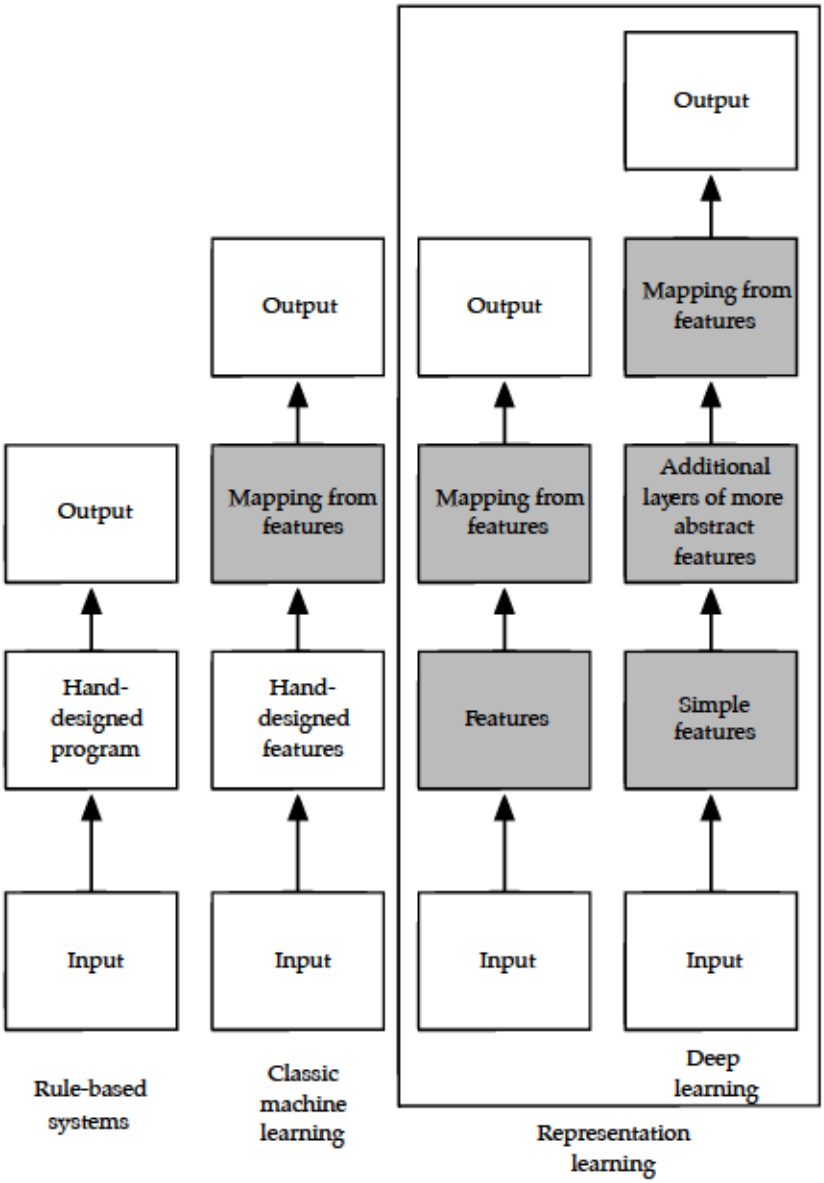



Figure 6.5: Overview of the machine learning techniques from the feature engineering point of view (Goodfellow et al., 2016). Boxes in grey show where the machine learning actually takes place.

words that in the process of tokenisation they will be represented by two different numbers, e.g. 124 and 52. Although there is a strong semantic relation between those words, after tokenisation that relation will be lost. There is no longer any relation between numbers 124 and 52 as their indices in the vocabulary. In other words, by changing the representation of the words, we have lost a big deal of valuable information that otherwise could improve the classification or extraction tasks.

The ideal situation would be to map words into a representation that would preserve their semantic meaning. In this case, we would not need to indicate *jefe* and *director* might sometimes have the similar meaning.

Two relatively recent works, word2vec and GloVe have achieved state-of-the-art in preserving semantic similarity by mapping words into high dimensional vector space (Mikolov, Corrado, et al., 2013; Pennington et al., 2014) in a completely unsupervised manner. The word embeddings of the aforementioned examples: *jefe* and *director* will be represented by similar vectors, close to each other in the vector space.

For the sake of financial event extraction, we have trained word2vec model based on the text coming from the entire news corpus. The training process is unsupervised, which means that apart from words and their contexts we do not need to provide any extra annotations. The training text consisted of single words but also frequent bi-grams and other n-grams representing named entities and other concepts from the financial ontology. We used the continuous bag-of-words (CBOW) model (Mikolov, Corrado, et al., 2013) to capture word semantics based on its syntactic context.

Essentially, the CBOW model is learning to predict a word given its surrounding context. For example in the following sentence we analyse the word "cat" in the context window of size 2:

"A $\overbrace{\text{large grey}}^{\text{left context}}$ **cat** $\overbrace{\text{was sleeping}}^{\text{right context}}$ on a rocking chair.

The context words for the word "cat" are: "large", "grey", "was", "sleeping". Those surrounding words are then used to train a shallow neural network in order to predict the word "cat". A similar analysis is per-

formed with every word in each sentence of the corpus. The objective function is defined as maximizing the probability:

$$\frac{1}{T}\sum_{i=1}^{V}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_i|w_{i+j})$$

where $T$ is training set size, $c$ is the context window size (amount of words to the left and to the right), and $w_i$ is the current word (and $..., w_{i-2}, w_{i-1}, w_{i+1}, w_{i+2}, ...$ is its context).

Let's define context words (the input for the training algorithm) as $w_I = x_1, x_2, ..., x_C$. The classification is then defined as a softmax function:

$$p(w_j|w_I) = p(w_j|x_1, x_2, ..., x_C) = y_j = \frac{\exp(\mathbf{v}'^T_{w_j}\mathbf{v}_{w_I})}{\sum_{j'=1}^{V}\exp(\mathbf{v}'^T_{w_{j'}}\mathbf{v}_{w_I})}$$

where $\mathbf{v}_w$ is the representation of the word $w$ at the input layer and $\mathbf{v}'_w$ is the representation of the word $w$ at the output layer (see Figure 6.6). As seen in the figure, the resulting network is a feed-forward neural network with one hidden layer.

The words in the input layer are represented as a *one-hot* vector of size $|V|$, where $V$ is the corpus vocabulary. The resulting matrix $\mathbf{W}$ is representing weights between the input layer and the hidden layer and the $\mathbf{W}'$ contains weights of connections between the hidden layer and the output layer. As the hidden layer size is $N$, then the matrix $\mathbf{W}$ is of size $N \times W$.

The one-hot vector representation is a sparse representation of the word $w_i$ that has a form of vector $x_i = \{0, 0, ..., 1, ..., 0\}$ with size of $|V|$ where 1 is at index $i$ (which is the same position as the position of word $w_i$ in the dictionary $V$). This representation has only symbolic mathematical meaning, as it is used to retrieve a dense represenation $\mathbf{v}_i$ of a word $w_i$ from the matrix $\mathbf{W}$ by simple multiplication:

$$\mathbf{v}^T_{w_i} = \mathbf{W}^T\mathbf{x}$$

that is: $\mathbf{v}^T_w$ is the row $i$ of $\mathbf{W}$.

The interpretation of the output layer is the vector $y$ of probabilities $p(w_j|w_1, w_2, ..., w_C)$, where the *output* vector representation $\mathbf{v}'_{w_j}$ of the word $w_j$ is the row j in $\mathbf{W}'$.

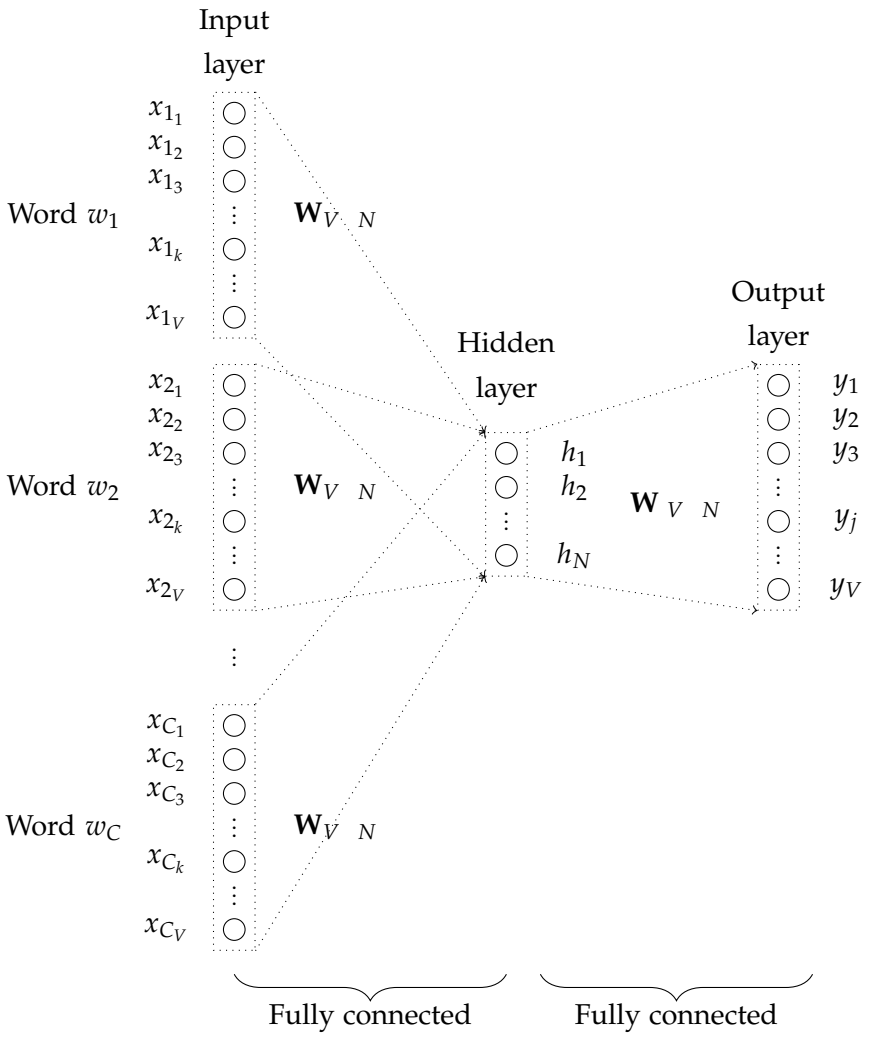



Figure 6.6: Continuous bag-of-words model as neural network with one hidden layer. Note that C is the total count of the context words (on the left side and on the right side), so $C \qquad 2c$

The neural network is using back propagation, stochastic gradient descent and hierarchical softmax methods for efficient calculation of gradients. The details has been explained in (Rong, 2014).

After processing the whole corpus we obtain dense representation of words in multidimensional vector space. As training parameters we used the continuous bag of words (CBOW) model, based on 5-surrounding words and 300-dimensional vector space. For the model computation we used gensim library[3] with fast word2vec implementation. The model's vocabulary consists of 1.718.075 words that appeared more than 3 times in the corpus.

The common property of obtained vectors is that words occurring in a similar context are close to each other in the vector space. In this sense, the word embedding is preserving the semantic similarity of those words. To visualise that fact, we created a small sample of words (and phrases) from different domains to assess how they cluster based on semantic similarity. Figure 6.7 shows the chosen subset of words mapped into a 2-dimensional space. As word vectors have 300 dimensions, we used t-distributed stochastic neighbour embedding (t-SNE) algorithm for dimensionality reduction (Maaten and Hinton, 2008). The t-SNE algorithm aims at preserving locality so that vectors which are close to each other in the high-dimensional vector space are also close in the low-dimensional projection[4].

After dimensionality reduction we applied density-based spacial clustering algorithm (DBSCAN) (Ester et al., 1996), setting parameters EPS=1.7 and allowing clusters of at least two point, to facilitate smaller cluster detection. Looking back into Figure 6.7 we observe clusters forming around words with very similar meaning and context. Each label colour represent different cluster, automatically assigned by DBSCAN.

What is noteworthy is that it does not only group words in the same context but can also distinguish between some very fine details. For

[3] API and documentation available at: `https://radimrehurek.com/gensim/models/word2vec.html`. Last accessed: 2016-10-11.

[4] t-SNE parameters that we used were: perplexity=15, theta=0.5 and no initial PCE preprocessing

instance, the cluster of companies (middle left) is divided between infrastructure and construction companies (in dark orange), banks (violet) and others (magenta). Even more surprising is the clean separation of companies: Pescanova, Gowex, Bankia and Rumasa (dark violet). Each of them was subject of a scandal of fraud or corruption. In the same way "Emilio Botín" (word bigram) did not form a cluster with "Rodrigo Rato" who is on the other far side of the banking cluster. He is rather closer to phrases like "Tarjetas black" or "Tarjetas opacas" (which are synonyms in this case) and point us to another scandal during his presidency in Bankia.

Another multi-cluster is related with politics (bottom centre), both Spanish and international. Spanish political parties form the first cluster (in purple), then right next to it a cluster of Spanish political leaders (green). The next cluster contains heads of state and government members. What is noteworthy is that Spanish Prime Minister is very close to this cluster, although he also belongs to one of Spanish politicians. Other sub-cluster of politicians are European level actors (such as Jean-Claude Juncker, Donald Tusk, Martin Schultz and Federica Mogherini). Next to it, Mario Draghi (President of the European Central Bank) and Christine Lagarde (Managing Director of the International Monetary Found) were classified together in a separate cluster, which seems appropriate due to their undeniable impact on the monetary policy and world economy.

We have also included words completely unrelated to politics or economy (in dark gray) and as expected the cluster has been pushed far from other topics.

Given that the word vectors preserve some similarities between different words it the question is if they also preserve relationships between them. Mikolov, K. Chen, et al. (2013) showed that semantic analogies between words are also represented in a vector space created by such embeddings. By performing algebraic operation on vectors it is possible to understand and discover some analogies between word vectors (see Figure 6.8)

The example given above can we written as algebraic operation on

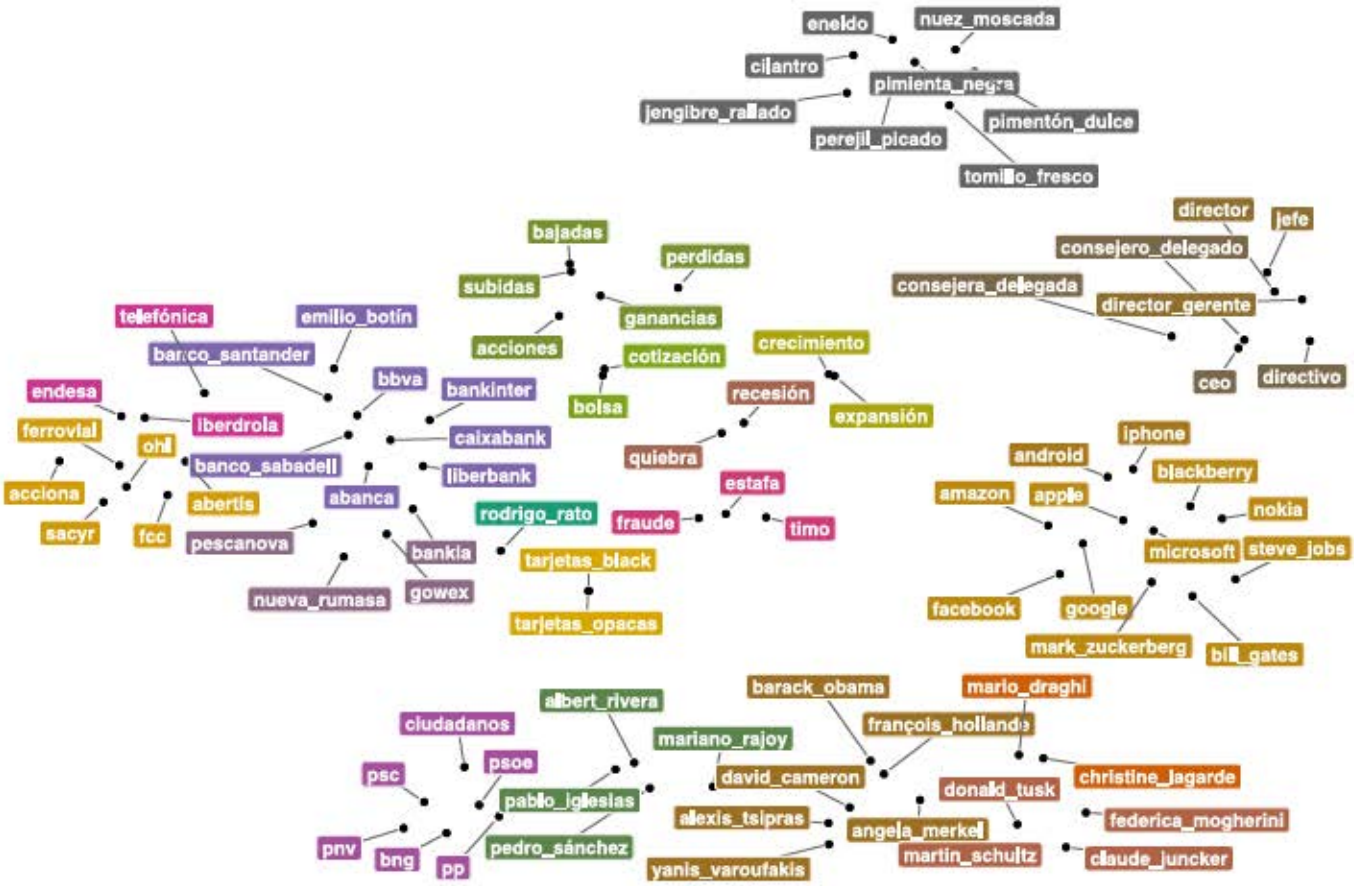



Figure 6.7: t-SNE projection of words vectors into a space of two-dimensions. Showing word clusters around similar topics and contexts. The chosen subset of words include some company names, key people and other words. Label colours were assigned automatically by the DBSCAN algorithm.

vector offsets:

$$queen \approx king - man + woman$$

which can be read as: "man is to king as woman to *queen*". The actual operation on the vector might not find the direct answer (e.g. a vector for word "queen") therefore rather a cosine similarity metric is used to find the nearest matching vector. If we define vector similarity in terms of highest value of cosine angle $\theta$ between word vectors $w_1$ and $w_2$:

$$\text{similarity}(w_1, w_2) = \cos\theta = \frac{w_1 \cdot w_2}{|w_1||w_2|}$$

Then finding nearest vector for analogy questions is to find $a'$:

$$\max(\cos(a', a - b + b'))$$

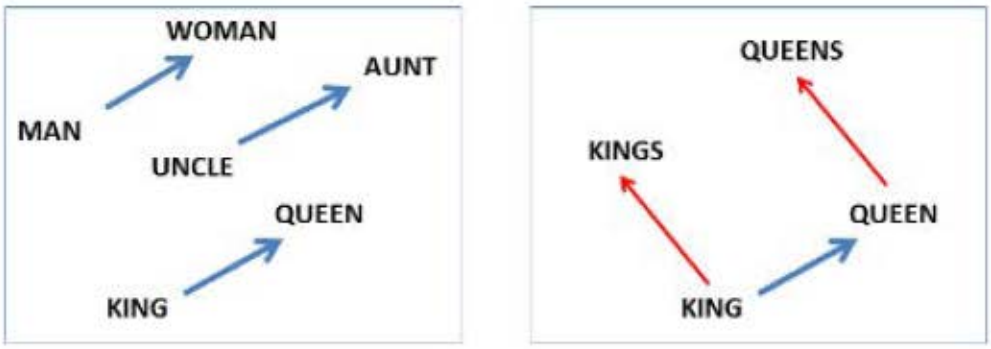



Figure 6.8: Projections of word vectors for gender analogy (Mikolov, W.-t. Yih, et al., 2013).

where words $a$, $a'$, $b$ and $b'$ are such as in the example above ($a$ – king, $a'$ – queen, $b$ – man and $b'$ – woman). Note that $\cdot$ symbol denotes a dot product of word vectors. For practical reasons if all vectors are normalized then the search is reduced to maximising the dot product: $a' \cdot (a - b + b')$.

We performed a similar search for in the news corpus model. Table 6.2 shows some examples for the word analogy questions. For general knowledge most of the answers were correct, however some less used words analogies might get a first closest match wrong. In the overview we provide also a few erroneous examples marked with an asterisk.

At this point, we see that the model is performing very well on word analogy task. In our case, it is not straightforward to objectively and quantitatively evaluate the model. For instance, Google published its word2vec model[5] trained on Wikipedia data along with a set of 20,000 queries with examples covering semantic and syntactic aspects of the model. But the same test is unsuitable for evaluating our news corpus due to different language and also slightly different scope of both corpora.

[5] The file can be obtained from `https://storage.googleapis.com/google-code-archive-source/v2/code.google.com/word2vec/source-archive.zip`, the evaluation file is: questions-words.txt. Last accessed: 2016-10-11.

Table 6.2: Word analogy examples from our trained embeddings model. The query result is in bold. Wrong answers are marked with an asterisk. For some cases we show second closest vector.

| *b* – *b* | *a* – *a* |
|---|---|
| chica – chico | mujer – **hombre** |
| chica – chico | abuela – **abuelo** |
| chica – chico | jefa – **jefe** |
| chica – chico | directora – **director** |
| chica – chico | reina – **príncipe*, rey** |
| chica – chico | informatico – **computadora*** |
| Pedro Sanchéz – Mariano Rajoy | PSOE – **PP** |
| Ciudadanos – PSOE | Albert Rivera – **Pedro Sanchéz** |
| Francisco González-Rodríguez – Emilio Botín | BBVA – **Banco Santander** |
| Banco Santander – Telefónica | Emilio Botín – **César Alierta** |
| Banco Santander – Bankia | Emilio Botín – **Miguel Blesa, Rodrigo Rato** |
| Banco Santander – El Corte Inglés | Emilio Botín – **Isidoro Álvarez** |
| Banco Santander – Gowex | Emilio Botín – **Jenaro García** |
| Banco Santander – Apple | Emilio Botín – **Tim Cook** |
| Google – Apple | Android – **iPhone** |
| Microsoft – Apple | Windows – **Windows Phone*, iOS** |
| Polonia – Portugal | Varsovia – **Lisboa** |
| Polonia – Francia | Varsovia – **París** |
| Polonia – Reino Unido | Varsovia – **Londres** |
| EE.UU. – Europa | dólar – **euro** |
| Berlin – Madrid | Alemania – **Portugal*, España** |
| negativo – positivo | decrecimiento – **crecimiento** |
| negativo – positivo | bajada – **subida** |
| negativo – positivo | empleo – **paro** |
| negativo – positivo | recortes salariales – **ajustes salariales** |
| positivo – negativo | luz – **electricidad*, oscuridad** |
| positivo – negativo | barato – **caro** |
| positivo – negativo | rápido – **lento** |

It is worth noting is that the word analogy is a result of semantics captured in the news articles, and is specific to this kind of source. The astonishing accuracy of analogy questions might be explained by the fact that the vector space captures a semantic representation of words accurately and such relationships naturally have their relationship-specific vector offsets in the vector space (Mikolov, W.-t. Yih, et al., 2013). Other explanation points to the fact that such relational similarity is captured by a group of highly indicative and rather rare words that are shared across different aspects (i.e. intersections of vectors $a$ and $b$ that are sharing the same aspect) and might be interpreted therefore as simply balancing words similarity (Levy and Goldberg, 2014).

Capturing word semantics is a crucial step, as we can now represent whole sentences with an array of word vectors and keep the meaning of each word in a phrase. This representation will be used in the next step, where we use deep neural networks in order to classify phrases for event detection.

## 6.5 Events classification with Convolutional Neural Networks

Our approach to classifying financial events is based on the state-of-the-art deep neural networks. Among many deep network architectures used in Natural Language Processing, the most effective and powerful are Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). Both provide a sound basis for creating even more complex designs, which can be observed nowadays in an exploding number of novel deep neural networks architectures. In this work we will focus on Convolutional Neural Networks for two reasons: (i) their good results in relation extraction tasks (Nguyen and Grishman, 2015) and (ii) relatively stable training and hyper-parameter optimisation comparing to RNN.

What makes Convolutional Neural Networks different from ordinary Neural Networks is the convolution step (LeCun, Bottou, et al., 1998), which aims at extracting features from a smaller window of input data. Originally convolution was aimed at extracting features from images by looking into smaller regions. This approach, however, proved to be successful in Natural Language Processing (Kim, 2014). The convolution operation is using a *filter* (or *kernel*) that slides over the input data and produces a *Feature Map* (or *Activation Map*). There can be many filters and each one trained to detect a different set of features. In the context of text classification, the convolution operation can be described as looking into hidden features represented by n-grams. Thus, the filter size describes the length of the n-gram. Moreover, those convolution steps can be stacked on top of each other in order to extract new different features at further layers.

After the convolution step, the feature map is passed through the activation function that introduces non-linearity, i.e. tanh, logistic sigmoid or ReLU. The features are further reduced through the pooling layer which reduces the number of features (dimensions) and keeps the most important (e.g. strongest) ones. The result of pooling is concatenated into one feature vector representing results of all filters. The last stage

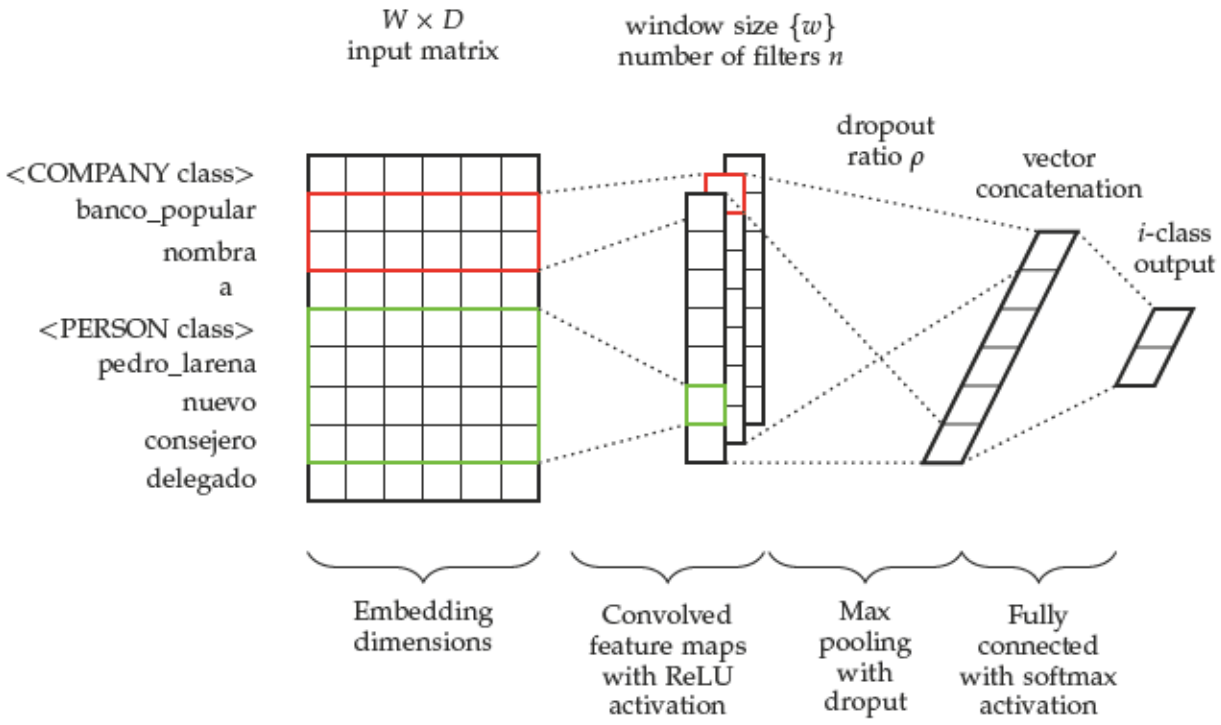



Figure 6.9: The architecture of the CNN network for event classification.

is the classification, that uses a fully connected layer that maps output features into concrete classes based on the training data, thus the output layer maps directly to the output classes.

Our approach to the event classification is based on the fact that we already use ontology resources and perform e.g. ontology-based NER on the textual input in order to identify relevant companies and actors. We extend the current CNN state-of-the-art approach to include an additional feature to the input data that are the ontology class embeddings.

Figure 6.9 shows the architecture of the CNN classifier. On the left side, we show the input sentence as it is prepared for the classifier. Before classification, the text is normalised to lower case with all special characters removed. Note that we are using bi-grams for some named entities (e.g. "Banco Popular", "Pedro Larena") as we already trained word embeddings with a list of popular bi-grams. If a specific bi-gram is not in the lookup list of word vectors, we simply treat each word separately. We include entity annotations by placing in front of each named

entity a "marker" vector (or "ontology class vector") to denote the class of the following entity. In the example given above those are: `<COMPANY class>` and `<PERSON class>`. This is based on the ontology-based NER annotations that happen before the text is actually fed to the classifier (see Section 6.3). The choice of ontology class vectors is arbitrary. The dimension values are generated only once, using uniform distribution with non-zero mean and small standard deviation.

Before feeding the input to the neural network, each word is mapped to its dense vector from the word embeddings lookup table (see Section 6.4). This results in an input becoming a matrix of $W \times D$, where $W$ is the max sentence size and $D$ is the dimensionality of the embedding vectors. Note that CNN requires all inputs to be of the same length. In case that the sentence is shorter, the remaining rows are padded with zero mean random vectors. For the input, we use only one word embeddings matrix (input *channel*), as there is no conclusive improvement in using multi-channel architectures.

For the sentence classification task we train binary classifiers, therefore the final layer consists of two nodes (positive/negative). In the case of relation extraction, we follow the approach of Nguyen and Grishman (2015), where we construct the input by appending the position embedding to the input vectors so that the input matrix is of size $W \times D + D_m$, where the $D_m$ is the length of position embeddings.

The next section provides more details on training and hyperparameters setup.

## 6.6 *Comparing classical machine learning and neural network approach*

We performed various classification tests with different classification scenarios with the purpose of improving financial event extraction. The comparison includes classical approach, deep neural networks approach and our approach combining deep neural networks with semantic taxonomy. As stated before, we perform the classification on the sentence

level, so in order to successfully classify sentence, all features, such as named entities need to be present, e.g. company name or person name. This is reflected in corpora of financial events for training. Therefore the fact that an event or relation is not contained in one sentence is not taken into account when evaluating corpus and classifiers.

### 6.6.1 *Classical machine learning approach setup*

As the bottom line classifier we use the Conditional Random Fields (CRF) algorithm with implementation and features described in (Surdeanu et al., 2011). The CRF also requires text preprocessing in order to extract various linguistic features. We setup the NLP pipeline (see Section 6.3) in order to provide all necessary input features. We evaluate CRF classifier in three different setups:

- CRF-Basic classifier is using CoreNLP: POS tagger, lemmatizer and parser and is not using NER in the process of classification.
- CRF-NER includes Spanish models for NER using the CoreNLP Stanford Named Entity Recognizer models[6] (Finkel et al., 2005) plus the ontology-based NER (see Section 5 for annotating companies).
- CRF-All is using the same setup as CRF-NER plus all the features for relation extraction from (Björne et al., 2011) that gives a comprehensive list of features to be included in the classification process.

### 6.6.2 *"Deep" Neural Network approach*

We evaluate the Convolutional Neural Network classifier in the following two variations:

- The CNN-Emb classifier is using Convolutional Neural Network with pre-trained word embeddings based on our corpus (see Section 6.4). The architecture of the classifier is described in the Section 6.5. The only difference is that this classifier is not using ontology class vectors.

[6] CoreNLP Models are from version 3.7.0, released on 31 Oct 2016.

- The CNN-Sem is based on CNN-Emb, but is using additional information based on ontology-based concept annotation (see Section 6.5). For NER tagging we used the same approach as for CRF-NER (see Section 6.6.1).

### 6.6.3 *Hyperparameters and training details of CNN classifiers*

We train both CNN classifiers with same the following parameters:

- for word embeddings (CNN-Emb, CNN-Sem) we use vector size of 300 dimensions (as explained in Section 6.4),
- the number of filters is 100,
- we use filter lengths of: 3, 4, 5,
- the dropout rate ($\rho$) is 0.6,
- learning rate of 0.001,
- the learning rate decay ratio is 0.7 every 16 epochs
- the batch size is 50,
- we run training for approximately 50 epochs.
- for each classification we divide the dataset set into a training set and evaluation set, where the evaluation set size is 10% of the whole dataset,
- for each classifier we perform a 10-fold cross-validation
- the final result is the average of each fold from the cross-validation,
- for the neural network training and evaluation we used `mxnet` open-source deep learning framework [7].

The choice of the parameters above was based on the hyperparameter optimisation through the random search (Bergstra and Bengio, 2012). A more detailed explanation of the meaning of each parameter can be

[7] The software is available at: `http://mxnet.io/`

found in (Wu, 2017). The same set of parameters proved optimal for both configurations (CNN-Emb and CNN-Sem). Figure 6.10 shows the training progress for CNN-Emb and CNN-Sem classifiers for the parameters above. We can observe that the error rate (loss) is lower for the CNN-Sem classifier (so at the same time the accuracy is higher).

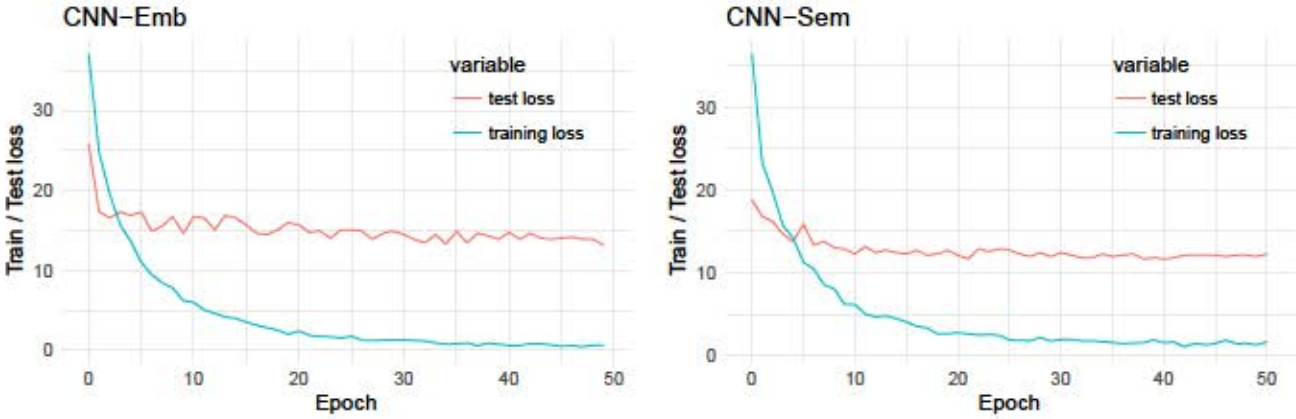



Figure 6.10: Training the CNN-Emb (left) an CNN-Sem (right) classifiers based on the given hyperparameters. The CNN-Sem shows lower test loss. The lines are averaged over all cross-validation folds.

### 6.6.4 *Classification results*

Table 6.3 shows an overview of the results for all evaluated classification configurations after 10-fold cross-validation. We can observe that the baseline classifier improves drastically when NER annotations are present, and even more when we use all possible (hand-crafted) features. On the other hand, the CNN-Emb classifier is significantly better that the bottom line classifiers even without any feature engineering. The CNN-Sem classifier, which uses ontology class annotations further improves the classification results and provides better performance in both: precision and recall measures.

Table 6.3: Results of different classification scenarios for single binary relation extraction.

| Classifier | Precision | Recall | F1 |
|---|---|---|---|
| CRF-Basic | 0.818 | 0.290 | 0.428 |
| CRF-NER | 0.797 | 0.852 | 0.823 |
| CRF-All | 0.865 | 0.816 | 0.840 |
| CNN-Emb | 0.919 | 0.891 | 0.904 |
| **CNN-Sem** | **0.943** | **0.896** | **0.919** |

Our results show that using deep neural networks can improve the process of information extraction. Using CNN-Emb with pre-trained word vectors has improved the binary relation extraction for our financial news dataset. This confirms hypothesis $H_4$ as evaluated against our corpus of Spanish news. Apart from Named Entities annotations, the CNN scenario did not require additional hand-engineered features in order provide better classifier, as opposed to the CRF-* scenarios.

The result for CNN-Sem classifier confirms hypothesis $H_2$, by providing an improvement over CNN-Emb (and all previous classifiers). The use of ontology-based annotations (as described in Section 5) can further improve the information extraction in financial domain, as presented in this study.

## 6.7 *Conclusions*

"Financial Event Extraction" step is one of the fundamental pieces of the whole investigation process. This is where the connection is being made between unstructured texts from the "Data acquisition" stage and

the semantic knowledge base envisaged in the "Semantic modelling" section. In this sense, the event extraction aims at classifying events and representing such occurrences in a semantic way, using models and taxonomies from Chapter 5. The aim to extract what we called *high-level* features: financial events and relations.

In this chapter, we focused on the Natural Language Processing aspects of the Decision Support System. We showed most important steps that form the NLP pipeline of the ALFREDO DSS, how the corpus for event extraction is being constructed and then we analysed the latest findings that can improve the overall Information Extraction process.

We showed that the dense vector word representation can preserve semantics, what traditional tokenisation does not. This representation became the input to the event classifier, where we employ our novel method based on artificial neural networks and ontology annotations in the text. We also peeked into traditional approaches to event extraction and sentence classification in order to compare with our approach. The conclusion is that in our setting they perform worse than our classifier. This observation allowed us to validate two research hypotheses: H2 and H4 (see Chapter 3)

This chapter paves the way toward the actual decision-making process. At this stage, we have almost all the pieces at had, except for one: the decision support model. We will discuss it in the next chapter.

# 7 *Decision-making based on unstructured data*

The purpose and importance of the previous chapters was to describe fundamental "building blocks" and detail the process to build a Decision Support System in the context of Business Intelligence. Those sections were paving the way for a bigger system that brings all previously described pieces together. This narrative has been brought in a bottom-up fashion: beginning from smaller, but fundamental pieces and joining them together to form the big picture. As we progressed through the previous sections, we were actually carrying out the investigation process devised in Chapter 3.

This chapter is the culmination of previous sections and brings together all the fragments into a coherent and complete architecture. The realisation of this architecture is the ALFREDO Decision Support System. We detail how previously described elements are working together towards the main goal which is the support in the financial decision-making process.

In the following sections, we: (i) describe the architecture of the ALFREDO DSS and the role of each main components, and (ii) define the hierarchical decision support model.

## 7.1 *Architecture for the Decision Support System*

Im this section we construct a complete architecture for analysis of news texts in order to provide a complete decision support system. The aim is to encompass all aspects of the methodological approach (see Section 3), describe stages of the data processing and model creation (as detailed in previous chapters), identify main components and depict how they work together in order to support the decision-making process.

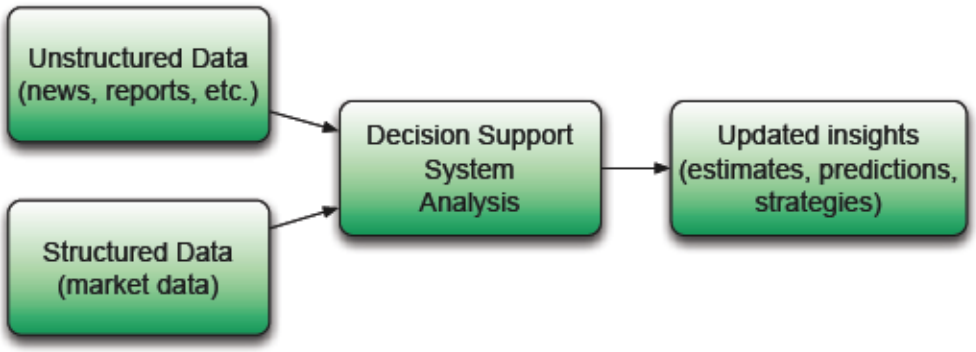



Figure 7.1: Overview of the news analytics for decision making.

The general idea for the news analytics is shown if the Figure 7.1. The DSS is fed with two principal kinds of data: structured (stock price time series, market data, etc.) and unstructured (textual sources, such as news). Both sources provide a stream of up-to-date information that updates our perception of the state of the world. Analysing that stream of data is crucial in order to properly react to the changing environment. Applying more sophisticated predictive analytics can further help to develop new insights, update investment strategies or make predictions about some unrolling events.

While traditional investment analysis relies mostly only on the structured data, the unstructured data can provide new insights to the overall decision making process. In the end this is what also drives the decisions of many investors. However, incorporating unstructured data is much harder, as it involves applying various domains of Artificial Intelligence such as Natural Language Processing, Machine Learning, etc. Overcom-

ing these difficulties can be rewarding, as it will make the whole process automatic. This can improve the overall decision making by widening the limits of investment data processing in a process called Intelligence Amplification (Leinweber and Sisk, 2012; Leinweber, 2009).

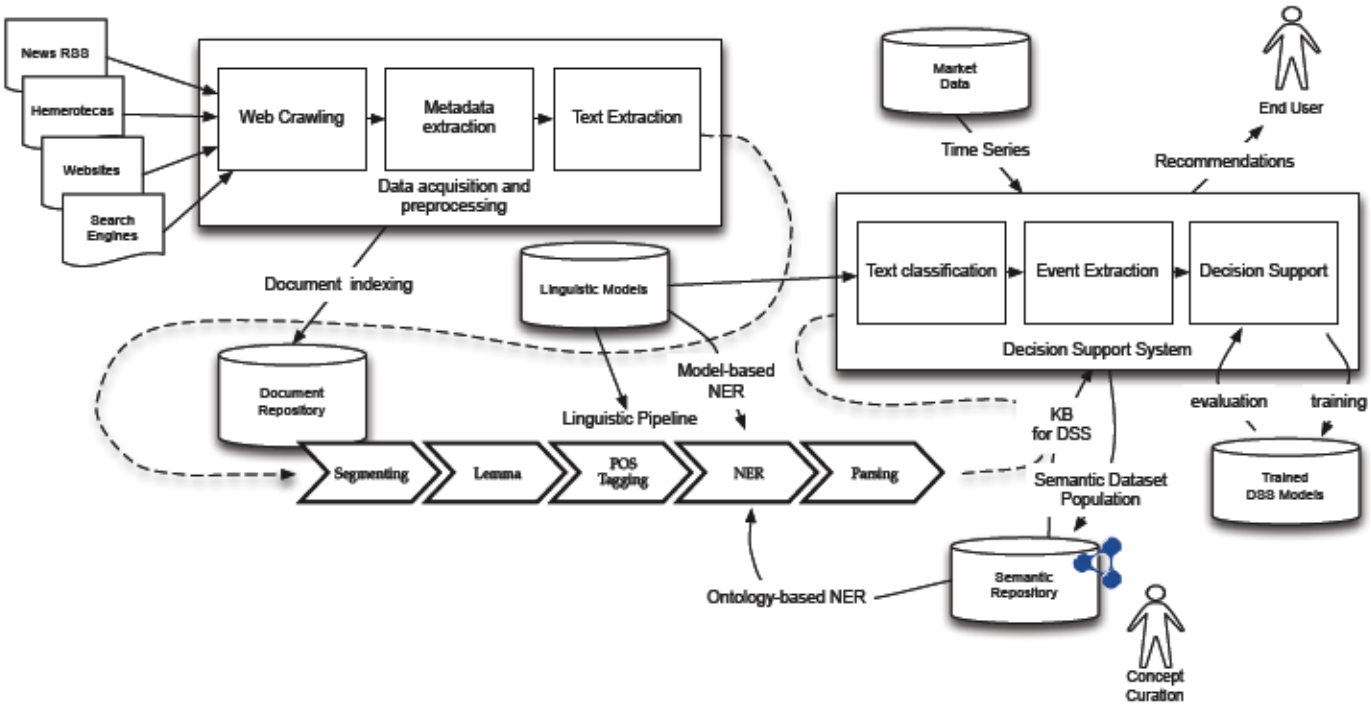



Figure 7.2: Overview of the ALFREDO Decision Support System architecture.

In the Figure 7.2 we introduce the architecture of the ALFREDO system. The system consist of the following main building blocks:

- Data Acquisition and Preprocessing — the main source of unstructured data comes from various external news sources. The text must be constantly acquired in order to provide timely data stream. After document acquiring, the text cleaning and metadata extraction steps are applied (as described in Chapter 4). Later, when the whole process is done, the document (along with its metadata) is indexed in a document repository. ALFREDO uses Apache Solr[1] for full-text index-

[1] Apache Solr is a popular open-source document repository with advanced search capabilities based on Lucene text indexing engine. The software can be obtained from the project website: http://lucene.apache.org/solr/

ing and faceted query search for further document retrieval based on document metadata.

- Linguistic (NLP) Pipeline — The role of linguistic pipeline is to analyse news text and extract new features that will further help spot relevant texts. Those features are syntactic and sentence-based (as opposed to the metadata from the previous step, which are document-based). We perform text segmenting (sentence splitting), lemmatisation, part-of-speech tagging, named entity recognition and sentence parsing. For NLP pipeline we use Apache UIMA framework (Ferrucci and Lally, 2004) with UIMA annotators and several extensions based on uimaFIT and DKPro software components (Eckart de Castilho and Gurevych, 2014b). It is important to mention that most of the steps are using linguistic models, while NER is using a hybrid approach: a trained linguistic model and ontology based concept extraction. While the ontology-based NER is used to tag entities that we control, such as companies and relate them with concrete ontology concepts, we also want to capture entities that are not yet in our ontology, for instance people names. For this reason we use hybrid NER and use both sources of annotations (see Section 6.1 for details.)

- Decision Support System — This is the final step of news processing, and also the most important one. So far we extracted text metadata and interesting syntactic features, we also tagged entities and linked them to our ontology. Now we are performing text classification in order to extract relevant financial events. In the text classification stage we aim at financial event classification on the sentence level (see Chapter 6 for more details). In the next step we perform extraction of additional information (such as entities involved) and produce semantic representation of newly extracted high-level features (see Chapter 5 for details). Semantic triples are stored in the Semantic Repository. We use RDF4J triplestore[2] (previously known as Sesame) to store and query extracted events. Based on news-extracted information, DSS

[2] RDF4J, previously known as Sesame is a popular RDF triplestore for storing querying and inferencing over semantic data

module provides decision support models. Those models are trained based on previously processed (historical) semantic data and external market data (time series) by the means of the machine learning algorithm (see Section 8.2). Recommendations are made by using our trained decision models to classifying newly incoming data.

- Linguistic Models for syntactic analysis — When performing NLP processing of news articles we use linguistic models for nearly all tasks. In the Linguistic Pipeline we use commonly available linguistic models from the Stanford CoreNLP toolkit (Manning et al., 2014). In the DSS we use our own trained models for financial event extraction, developed within this thesis.

- Semantic Knowledgebase — This is the main repository storing the extracted knowledge and used by DSS for both training and decision-making process. Apart from extracted knowledge it also contains the financial ontology (vocabularies and taxonomies). The part of the ontology describing named entities (people, business organisations, places) is also used for ontology-based NER in the Linguistic Pipeline. It convenient to assume that the semantic knowledgebase can be populated automatically without any human intervention. The fact is, that ontologies need curation and it is important that most relevant taxonomies are up-to-date. While DSS-extracted facts can be erroneous sometimes, it is important that the ground truth, core ontology be accurate as much as possible. In our case, the critical part is named entities taxonomy which should be periodically revised in order to keep the system accurate.

## 7.2 *Model for decision support*

The core idea of the decision support model is the gradual analysis of data, by starting from raw text and build the way up from low-level features to more abstract relations and features. The idea of such hierarchical model is presented in Figure 7.3.

This hierarchical model explains the gradual process of text process-

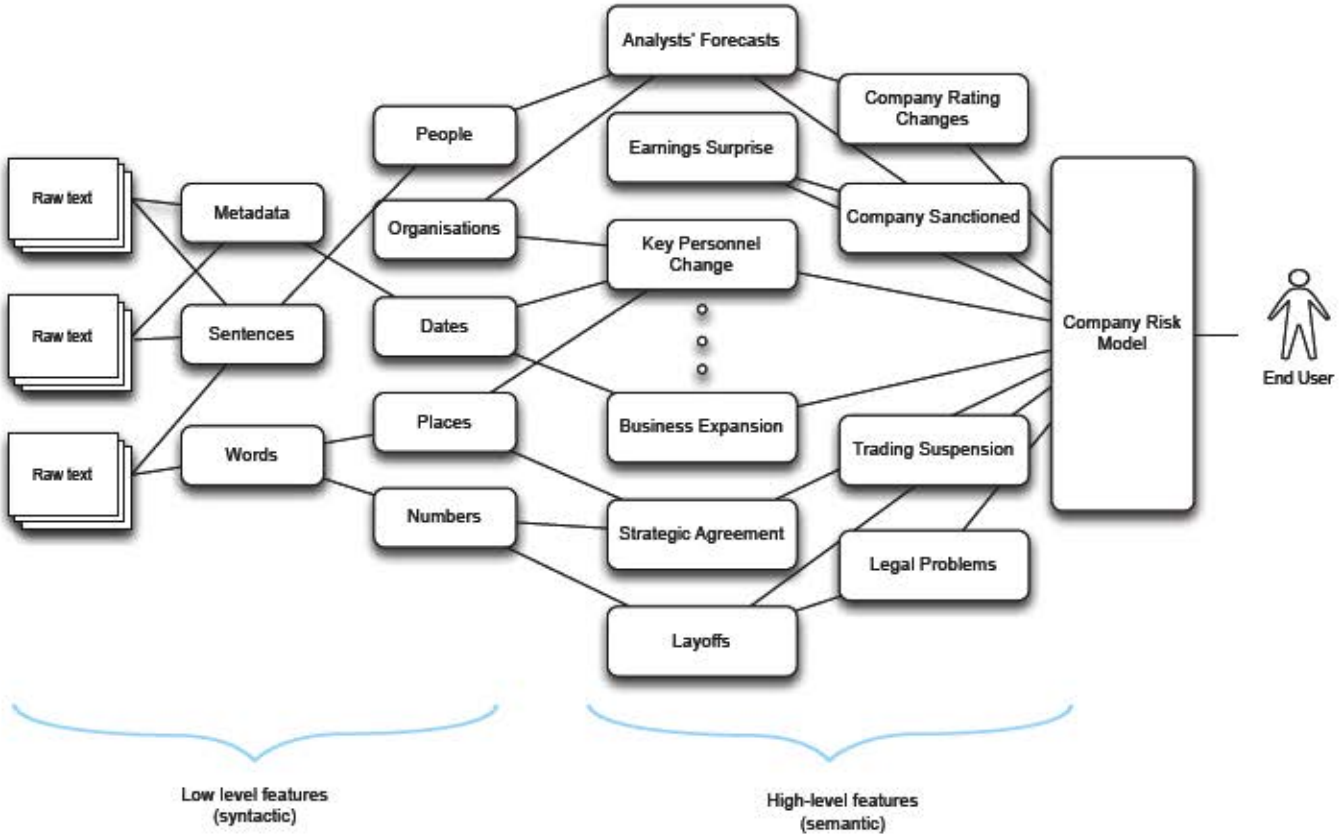



Figure 7.3: Hierarchical Model for Financial Decision Support

ing and enrichment that starts with raw news text, that is a stream of characters. Then the extraction of *low-level features* takes place that separates text into words, sentences and additional metadata (on a document level). This is where the main Natural Language Processing takes place. The NLP pipeline further generates additional NER annotations, such as key people, company mentions, dates, places, numbers etc. At this point, the system starts linking extracted features with their semantic concepts. This results in a stream of low-level features, that provide atomic primitives (in a form of mentions) that are later used to support the extraction of the *high-level features*. By this term we understand financial events that are most valuable for financial analysts: events that occur in time and meaningful relations between financial entities, people and other semantic concepts.

At this point, the DSS lifts relevant concepts (entities, events and rela-

tions) from syntactic form to the semantic representation. The extraction of events that works on semantically enriched text (but still: a text) produces semantic triples that are stored in a triplestore, according to the semantic model described in Chapter 5. This information is then used to train a DSS model (more on that in Section 8.2).

While it is difficult to know *a priori* what kind of events can have an impact on the company value, we let AI machine learning algorithm to learn this from the past data. This is contrary to other approaches, e.g. in (Nuij et al., 2014) we don not adjust weights associated with each event, but rather let the machine learning algorithm learn it. For this we use past news from the corpus as described in Chapter 4.2 and combine this information with market data (Sprenger and Welpe, 2011) in order to train our DSS model (more on that in Section 8.2).

## 7.3 Conclusions

"Decision-making based on unstructured data" introduces a theoretical model for the ALFREDO DSS. It gathers all the pieces described in the previous steps into one coherent architecture for the decision-making process. We detail all the pieces and explain how they work together towards the final goal that is providing financial decision support to the end-user. As the whole process is data-driven, the data flow has been carefully explained: how the raw unstructured data is piece by piece transformed into semantic knowledge and the used as an input for the decision models.

We also explained the role of each piece of information extracted on the way: *low-level features* providing base metadata, NER process for spotting relevant entities in the stream of documents, and *high-level features* for the actual decision-making process.

At this point, the design of the DSS is complete. We are now ready to perform the evaluation of the overall approach against the real-world data.

# 8 Evaluation

The essential characteristics of Decision Support Systems is their impact on the surrounding reality through decisions and recommendations they produce that can solve real-world problems. Therefore DSSs need real realistic setup and real-world data in order to prove their usefulness.

There are many approaches to validate the usefulness of a Decision Support System. They can be generalised into two groups. The first are qualitative techniques, where experts' feedback is taken into account in order to assess and validate results produced by DSS. On the other hand, there are quantitative methods, where empirical observations are analysed and compared in order to draw conclusions and validate research hypotheses. In this case, the result is more objective as it relies solely on data, but it also requires a rigorous analysis to ensure correctness. We will focus on the latter method as a way of evaluating this thesis' work and assessing the viability of the proposed DSS.

The remain part of this section consist of: (i) description of evaluation context and the backtesting details, (ii) DSS training on the real market data, (iii) evaluation scenarios, (iv) evaluation results, (v) discussion of the validity of the evaluation results.

The whole evaluation process is performed in the real-world setting in the context of the Spanish market and IBEX35 companies.

## 8.1 Backtesting evaluation context

The validation is performed using the actual stock prices of Spanish companies traded on the Madrid Stock Exchange. Recommendations produced by ALFREDO DSS are evaluated in a process called *backtesting*. This essentially means running a strategy against past data in order to evaluate how it behaves and what is its performance ratio. When we backtest a strategy, we divide past data into two separate periods: one for training and the other for evaluation. It is important that they do not overlap so that the training classifier never sees any evaluation data beforehand nor any of the training data is seen in the evaluation.

For the validation task, we divided the corpus data into two periods: the training is from 2014-01-01 to 2015-12-31, and evaluation from 2016-01-01 until 2016-05-31.

The length of the training period is due to the fact that we need to establish enough knowledge on past facts (events) to be able to successfully train our decision model. Figure 8.1 shows the training and evaluation periods as an overlay of the IBEX35 index. The training period (in green) covers 2 years, while the validation is the remaining 5 months (in orange). Note that the validation period is quite volatile and returns on most IBEX35 companies were negative.

When including the market data, we are evaluating price evolution using daily time series. This price development is analysed in terms of relative rather than absolute values. This is because we are more interested in the *change* of price values (i.e. rising or falling), not in the value itself. There are two main ways of calculating returns: (i) using percentage returns, (ii) using log returns. Both have their advantages in different contexts. As stated above, we will focus on price evolution in the longer time periods and we will compare it with other time series, therefore it is more suitable to use the percentage returns.

For this reason, we calculate daily returns, defined as a 1-day price change (percentage) in respect to the previous value:

$$R(t) = \frac{S(t) - S(t-1)}{S(t-1)}$$


Figure 8.1: The backtesting period divided into the training (green) and evaluation period (orange). Note that in the whole time frame, IBEX35 is losing about 8% points, half of which happens on the evaluation period.

where the $S(t)$ is the price on the day $t$.

The source market data are daily closing prices for each stock. If the previous day is not a trading day, then by $t-1$ we refer to the last trading day. The original time series data are *adjusted* for events such as dividends, stock splits, reverse stock splits or distributions. For instance, if a 1-to-2 stock split occurs, the price of a share drops two times. Such a sudden price gap could be potentially misleading and could be erroneously taken as an indicator of some important event. In such case, adjusting the price is to divide all historical prices before the split, so the time series is smooth again.

Obviously, the adjustment of price data changes the absolute historical prices, but has no effect on the relative values (e.g. returns). This is however desirable from the point of view of this evaluation, where we

look only at returns rather than absolute values.

## 8.2 DSS model training

Before training the model we need to prepare the training dataset that consists of relevant financial events that had an impact on the stocks in the past. This process is difficult to perform manually as it would require us to annotate a huge amount of documents. Another problem is the changing economic situation, for example, a dataset annotated before the Lehman Brothers[1] news can be very different than now. Some events can be incomparable, have a different impact, etc. Keeping such dataset up-to-date would require constant revisions and updates which in consequence would rise the effort of an already costly procedure.

Instead, we use *distant supervision* technique for creating the training set for DSS model training. This technique uses an *event study* to perform an empirical analysis of past data in order to find days (and related news) where some unusual event occurred. We repeat this for all IBEX35 companies, and for the whole period of training in order to create a complete training dataset.

The assumption behind this step is that the corporate news is driving price changes to some extend (Ryan and Taffler, 2004), therefore we aim to find news that is indicative of relevant economic events. We want to run this test for each company in order to spot those days and further extract relevant news that appear on that day and mention that certain company. The training dataset can be described as a set of documents $\{D_{i,t_e,o}\}$, where $D_{i,t_e,o}$ is a news document mentioning company $i$ on event day $t_e$ and with orientation $o$. The orientation $o$ indicates if an event has a positive or negative impact on the company $i$ stock price.

For event study we use the *market model* (MacKinlay, 1997) which builds upon the correlation between actual return of a company stock

[1] The Lehman Brother bankruptcy in 2008 started a chain of events that lead to the global financial crisis.

and expected return for this company:

$$AR_{i,t} = R_{i,t} - E(R_{m,t})$$

where $AR_{i,t}$ is the abnormal return for a company $i$ on day $t$, $R_{i,t}$ is the actual return and $E(R_{m,t})$ is the expected return given the absence of event. This expected return can be expressed as a relation to the reference market (e.g. market index):

$$E(R_{m,t}) = \alpha_i - \beta_i R_{m,t}$$

The $\alpha_i$ and $\beta_i$ parameters are defining the assumed constant and linear relationship between the company $i$ stock and the market. We will be estimating those parameters for each stock using ordinary least squares method (OLS or linear least squares), which is simply fitting a linear regression model that minimises the sum of square root error (Brown, S. J. and Warner, 1980). The parameters are estimated on the estimation window (as shown in Figure 8.2).

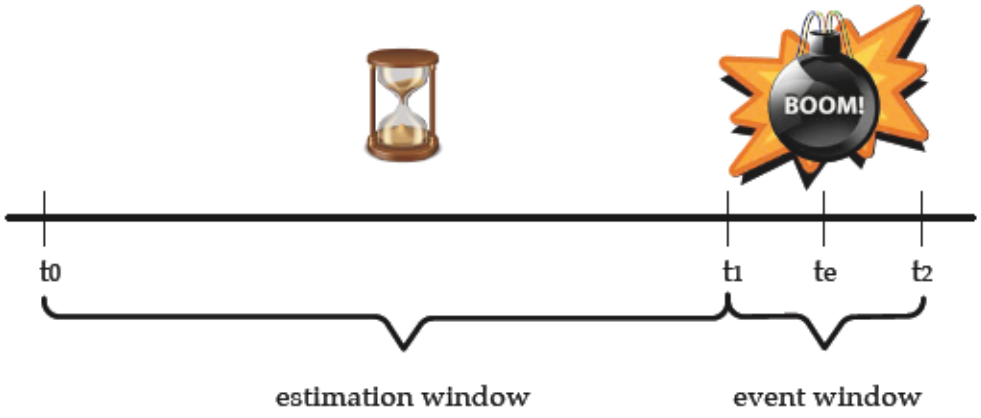



Figure 8.2: Event study model parameters.

For calculating abnormal returns ($CAR$) we are using aggregated measures over abnormal returns ($AR$), defined as:

$$CAR(t_1, t_2) = \sum_{t=t_1}^{t_2} AR_{i,t}$$

and, for cross-sectional cumulative average abnormal returns ($CAAR$):

$$CAAR(t_1, t_2) = \frac{1}{n} \sum_{t=t_1}^{t_2} CAR_{t_1,t_2}$$

To perform the event study we employ traditional parametric Cross-Sectional t-Test described in (Brown, S. J. and Warner, 1980, 1985). The test is defined that, under the null hypothesis, the cumulative average abnormal return is equal to zero:

$$H_0 : CAAR = 0$$

The statistic for this test is defined as:

$$T_c = \frac{CAAR(t_1, t_2)}{\hat{\sigma^2}_{CAAR(t_1,t_2)}}$$

where variance is estimated on cross-section of cumulative abnormal returns:

$$\hat{\sigma^2}_{CAAR(t_1,t_2)} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \Big( CAR_i(t_1, t_2) - CAAR_i(t_1, t_2) \Big)^2$$

The estimation period for calculating the cross-sectional standard deviation of companies' means is one full year before the backtesting. For instance, for checking the day of 5 January 2014 ($t_e$) we start the estimation period on 4 January 2013 ($t_0$). The event window is defined as a 3-day window (one day before and one after the event), so the event window start date $t_1$ is 4 January 2014.

We run the event study test for each day of the training period and for each company in the IBEX35 index.

We reject the null hypothesis (that is: we assume that a current day is an event) on t-Test $\alpha$ value below 0.01, which gives us the 99% confidence level. This helps us focus only on relevant events, with relatively large impact on the cumulative returns.

After creating the list of events for each company, we extract news from those days, but only mentioning the given company and run them through the ALFREDO pipeline. For instance, for day 5 October 2015, on which Banco Santander registers unusual return, we retrieve all news mentioning this company on this day. By whole day, we mean all news since last market closing hour until the market close on the event day. The market close time is typically around 17:35 CET after the closing auction is over.

After all, documents are processed and relevant events have been extracted we create the training set containing both: semantic features,

syntactic features and raw news text. We assign positive and negative labels depending on the orientation of a concrete event. If the return was positive that day, we assign 1 or 0 otherwise.

With this information, we train a machine learning model in order to learn the label 0 or 1 (i.e. the price response to an event) based on the set of the aforementioned input features. We evaluated several machine learning algorithms (gradient boosting, support vector machines and feed-forward neural networks), but the results were similar to all variants. We finally used the feed-forward neural network (with 28 hidden nodes, the learning rate of 0.05) and we obtained the evaluation loss of 0.216. This result is given from the purely informative reasons, as it does not yet provide any insight of the DSS performance yet. It is rather a final step towards the DSS model creation, and the actual evaluation is performed in the next chapter.

## *8.3 DSS model evaluation scenarios*

We perform the evaluation of the model based on results of ALFREDO recommendations that are made only by processing the news text. Recommendations produced by ALFREDO have a form of binary values (in literature it is also called a "signal") defining the recommended position that should be taken for a given company stock in a given day. The position can be either long, i.e. buying the stock, or short, which means selling a borrowed stock. The long position means that we expect prices to rise, therefore we acquire a stock. A short position is slightly more complicated. In this case, we expect the company value (and its corresponding stock price) to fall so we make an agreement to owe the stock for a given time. When the stock price fall we can buy the stock for a lower price and fulfil the agreement.

Those recommendations are further compared with the market data, comparing our position against the performance of each company stock value. As a result, we create a cumulative results table for each trading day. When analysing performance, we are simply following IBEX35 companies without any portfolio rebalancing nor management, as this

would go far beyond the scope of this thesis. We simply evaluate all companies, as if we were investing in each one of them at the same time.

We validate hypothesis H1 and H3 through various evaluation scenarios. In the first scenario assess the ALFREDO system by analysing all news and corresponding companies' results on the same day $T-0$ (read: T minus zero, which is effectively a test with a 0-days lag). Specifically, we are looking at the news from the whole event day $T$, making a prediction and looking at the closing price of the same day in order to check if the prediction is correct. This means that we want to know if the decision model (together with the whole underlying architecture) can actually identify relevant financial events and provide an accurate recommendation. By evaluating ALFREDO on $T-0$, we also implicitly assess if the event extraction pipeline can actually produce relevant knowledge. This scenario is designed to confirm or reject the hypothesis H3: by the means of providing evidence if the automatic process of model training and semantic facts extraction leads to concrete analytical capabilities.

Note that in $T-0$ we can not actually take any position (neither long nor short), as to do so we should have known the result the day before. It is still useful information though, as it provides insight into the current situation of a company at hand.

In the second evaluation scenario, we analyse the news on $x$ days before in order to take a position and close it in the future. We perform the test for $T-1$ (1-day lag), $T-2$ (2-days lag), $T-3$ (3-days lag) and $T-4$ (4-days lag). In this scenario, we assess if ALFREDO can actually provide any information that can help make future decisions. If the analysis of news text has any predictive power, this evaluation should give positive results in comparison to the market results. The result of those scenarios is important for validation of hypothesis H1.

We evaluate ALFREDO DSS with the most commonly used metrics to measure the impact of events on the stock prices evolution (Geva and Zahavi, 2014), those are: (i) cumulative returns and (ii) Sharpe ratio (Sharpe, 1994). As the performance reference, we use the IBEX35

index. Generally, if the result is better than the market result then we can consider it successful. We run the whole evaluation experiment on the testing period from 1 January 2016 until 31 May 2016 (5 months in total).

Cumulative returns are calculated as a sum of daily returns, based on ALFREDO recommendations. When a recommendation anticipates the actual market movement of a given company stock, the daily return for that company has a positive value. Otherwise, it is negative. The Sharpe ratio is defined as a ratio of reward and variability, which essentially examine the performance adjusted for its risk, and is defined as follows:

$$S_r = \frac{\overline{(R_a - R_f)}}{\sqrt{\sigma_{(R_a - R_f)}}}$$

where $R_a$ is the asset return and $R_f$ is the risk-free asset, which may be either a market index or government bonds (e.g. 3-months Treasury Bills, in Spanish: *Letras de Tesoro a tres meses*). Thus, the Sharpe ratio values excess returns but only when the investment decision does not involve excessive risk. In our case, in all calculations of Sharpe ratio we use the IBEX35 index as a risk-free asset reference.

## 8.4 Evaluation results

Table 8.1 shows a summary of the performance for each evaluation scenario run. First of all, looking at results of the test $T - 0$ we can see that the model actually learns very well how to deal with financial events. It can correctly recommend a profitable position 67% of the time, which results in 161% return across all companies, with Sharpe Ratio of 0.3 (in the period of 5 months). Given that the model learns only from the past knowledge this convalidates analytical capabilities behind ALFREDO processing pipeline.

Other important tests are the lagged backtesting scenarios (from $T - 1$ to $T - 5$) where we evaluate out recommendations against the future price movements. The following results from the table 8.1 show that only the $T - 1$ test yields positive returns. Tests $T - 2$ and $T - 3$ are very

close to the market results, and $T-4$ and $T-5$ are below the market performance.

| Test | Cumulative Return (average) | Sharpe Ratio (average) | Win (%) (average) |
|---|---|---|---|
| **T-0** | **1.61** | 0.30 | 0.67 |
| **T-1** | **0.09** | 0.02 | 0.57 |
| T-2 | -0.04 | -0.01 | 0.56 |
| **T-3** | **0.00** | 0.01 | 0.57 |
| T-4 | -0.07 | -0.03 | 0.57 |
| T-5 | -0.10 | -0.04 | 0.56 |
| IBEX35 | -0.03 | -0.01 | 0.52 |

Table 8.1: Performance evaluation results of the ALFREDO decision support system. Cumulative return of 1 means a 100% return. The last row shows the IBEX35 index results in the same period as a reference. Non-negative results are in bold.

The evolution of the cumulative returns of the test $T-1$ together with cumulative returns for the IBEX35 index is shown in the figure 8.3. The results are aggregated across all companies as if we were investing in every one of them. Apart from better cumulative returns, we can notice that the losses are less severe than on the original IBEX35 index.

Another overview evaluation scenarios result is presented in the Table 8.2. We show results on the per-company basis for $T-0$ and $T-1$ scenarios. In the last column, we show the total amount of news mentions for each company in the testing period. For details on the cumulative returns for each individual company, see Appendix D.

Table 8.2: Evaluation scenarios results per company.

| Company | $T$ 0 scenario | | | $T$ 1 scenario | | | |
|---|---|---|---|---|---|---|---|
| | Cumulative Return | Sharpe Ratio | Win % | Cumulative Return | Sharpe Ratio | Win % | Number of mentions |
| Abertis A | 0.25 | 0.15 | 0.57 | -0.16 | -0.10 | 0.47 | 2838 |
| Acerinox | 6.82 | 0.89 | 0.83 | -0.08 | -0.01 | 0.52 | 1475 |
| ACS | 0.75 | 0.27 | 0.63 | 0.05 | 0.03 | 0.54 | 3039 |
| Aena | 0.56 | 0.33 | 0.64 | 0.30 | 0.19 | 0.61 | 3487 |
| Amadeus | 0.17 | 0.13 | 0.70 | -0.05 | -0.03 | 0.55 | 993 |
| Arcelormittal | 20.57 | 0.76 | 0.76 | 1.59 | 0.21 | 0.56 | 3626 |
| Banco Popular | 3.32 | 0.36 | 0.67 | -0.06 | 0.01 | 0.52 | 2392 |
| Banco Sabadell | 1.29 | 0.28 | 0.65 | 0.24 | 0.08 | 0.45 | 6483 |
| Bankia | 0.96 | 0.24 | 0.53 | 0.18 | 0.07 | 0.51 | 17155 |
| Bankinter | 0.66 | 0.28 | 0.64 | 0.03 | 0.02 | 0.57 | 3808 |
| BBVA | 0.86 | 0.23 | 0.62 | 0.22 | 0.08 | 0.54 | 15459 |
| Cellnex | 0.48 | 0.28 | 0.66 | 0.19 | 0.13 | 0.59 | 1579 |
| Enagas | 0.09 | 0.20 | 0.84 | 0.01 | 0.04 | 0.83 | 122 |
| Endesa | 0.09 | 0.08 | 0.55 | 0.31 | 0.23 | 0.61 | 4018 |
| Ferrovial | 0.09 | 0.06 | 0.53 | -0.01 | 0.00 | 0.56 | 3210 |
| Gamesa | 0.93 | 0.23 | 0.71 | 0.08 | 0.04 | 0.51 | 4351 |
| Gas Natural | -0.10 | -0.05 | 0.47 | 0.27 | 0.15 | 0.58 | 2602 |
| Grifols | 0.50 | 0.35 | 0.73 | -0.05 | -0.04 | 0.58 | 1482 |
| IAG | 1.46 | 0.41 | 0.66 | -0.04 | -0.00 | 0.57 | 2263 |
| Iberdrola | 0.60 | 0.48 | 0.70 | -0.12 | -0.12 | 0.48 | 7959 |
| Inditex | 1.75 | 0.76 | 0.82 | -0.05 | -0.02 | 0.49 | 5310 |
| Indra A | 1.69 | 0.40 | 0.67 | 0.39 | 0.13 | 0.50 | 3475 |
| Mapfre | 1.45 | 0.39 | 0.66 | 0.19 | 0.08 | 0.54 | 2601 |
| Mediaset | 0.18 | 0.08 | 0.50 | 0.22 | 0.10 | 0.60 | 2429 |
| Meliá Hotels | 0.12 | 0.39 | 0.94 | -0.08 | -0.27 | 0.92 | 46 |
| R.E.C. | 0.05 | 0.05 | 0.50 | 0.10 | 0.11 | 0.56 | 1336 |
| Repsol | 1.86 | 0.36 | 0.69 | -0.36 | -0.12 | 0.49 | 8866 |
| Santander | 0.06 | 0.03 | 0.52 | -0.06 | -0.01 | 0.49 | 15946 |
| Tecnicas Reunidas | 1.85 | 0.35 | 0.68 | -0.36 | -0.12 | 0.49 | 1364 |
| Telefonica | 0.60 | 0.22 | 0.63 | -0.03 | -0.00 | 0.48 | 18233 |
| Viscofan | 0.09 | 0.21 | 0.92 | -0.05 | -0.10 | 0.90 | 156 |

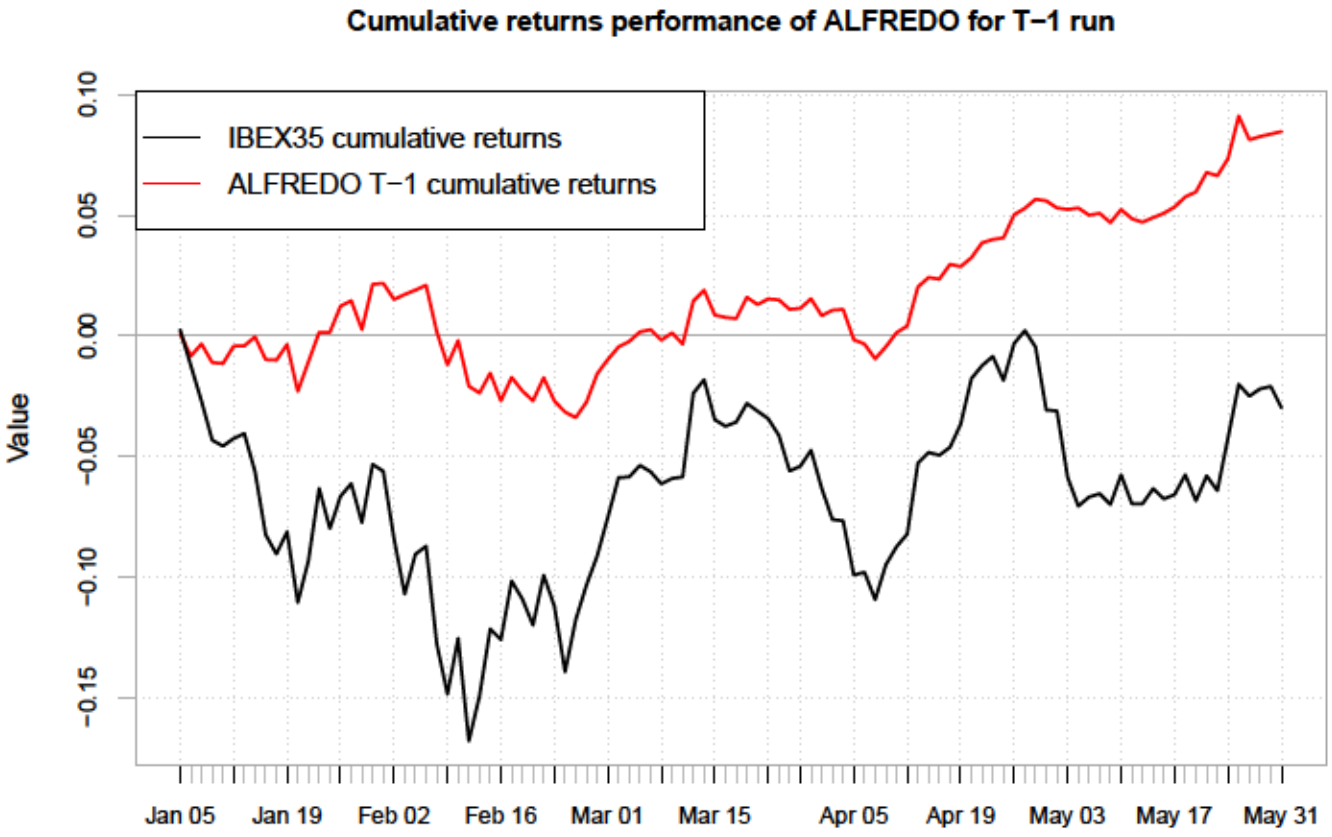



Figure 8.3: Comparison of cumulative returns between IBEX35 stock and ALFREDO $T-1$ evaluation scenario, as aggregated over all companies.

The result of the $T-0$ scenario confirms our hypothesis H3: ALFREDO is able to automatically learn through the distant-supervised learning from the past data in order to recognize and correctly classify the impact of real-world financial events 67% of time and resulting in a cumulative return of 1.61 and Sharpe ratio of 0.30.

It is noteworthy that while we use real-world data, we do not include some practical aspects of stock trading, such as: broker and exchange fees, payments for stock borrowing (when short-selling), the *slippage* effect (when the order is placed not at the right moment and results in an unfavourable price) or any other possible trading limitations[2]. Those aspects might be very important in practice when doing actual operations. However including them here is beyond the scope of this thesis.

## 8.5 *Validity of ALFREDO recommendations*

One important question that we have to ask ourselves is: is our DSS truly making better informed decisions or it is simply sheer luck? In other words: are we really making correct predictions or maybe we constructed a very sophisticated coin-flipping machine? It is rather difficult question to answer, especially regarding the movement of stock prices that is a result of a very financial complex system involving many actors. We can however test how our predictions compare to the random data, if they can be at least statistically distinguished from the noise. This is especially important in the context of lagged tests (i.e. T-1 and later). If we want to confirm hypothesis H1, we need to be sure that our result is not simply a matter of luck.

We create QQ-plots for each validation scenario in order to compare the distribution of ALFREDO recommendations with a random (uniform) distribution. This test shows how similar are both distributions by comparing quantiles of each one on the same plot. That is, each point's coordinates contains a corresponding quantile of uniform distribution ($x$) and distribution of ALFREDO positions ($y$). For similar distributions, those points are aligning on the $y = x$ line. Any other shape means that both distributions are of a different nature.

Figure 8.4 shows a QQ-plots of results from the ALFREDO backtesting scenarios. For scenarios $T - 0$ and $T - 1$ we observe that all the points

[2] For instance, in order to reduce volatility, short-selling was restricted by the Spanish national regulator (CNMV) from mid 2011 and until the beginning of 2013

are rather far from forming a $y \quad x$ line, which clearly shows that both distributions are different from a uniform (random) distribution. For those plots, the $y \quad x$ does not approximate well the trend of the data. In those cases, we can see that it is unlikely that those are random "lucky" strategies. On the other hand, in all remaining scenarios ($T \quad 2$, $T \quad 3$, $T \quad 4$ and $T \quad 5$) all points are trending around the $y \quad x$ line, suggesting that those scenarios are very likely producing just random "noise", even if the $T \quad 3$ scenario is slightly better in terms of returns.

Based on this conclusion, we can validate hypothesis H1, as we are able to improve the decision-making process by producing a better return than the market average, as represented by the IBEX35 index. We also show that in the case of the "predictive" scenario ($T \quad 1$), the nature of ALFREDO predictions are not a matter of luck, but a result of the sound decision model.

## *8.6 Conclusions*

The "Evaluation" is the final stage of the whole methodological approach that we defined at the beginning. This is where we evaluate our whole approach to the decision-making process that is represented by the ALFREDO DSS.

In order to perform a data-driven evaluation, the first step is to provide a training and testing datasets. This task is already quite difficult from two perspectives: (i) manually creating such dataset is laborious, (ii) the dataset should cover sufficiently long time span, making it infeasible with traditional techniques. Thus we performed an event study in order to detect significant events and from that create a news corpus to process through the ALFREDO DSS in order to train the decision model.

The evaluation was performed in multiple scenarios in order to assess: (i) the capability to learn from existing events and, most of all, (ii) anticipate future reactions to events. We showed that the DSS model

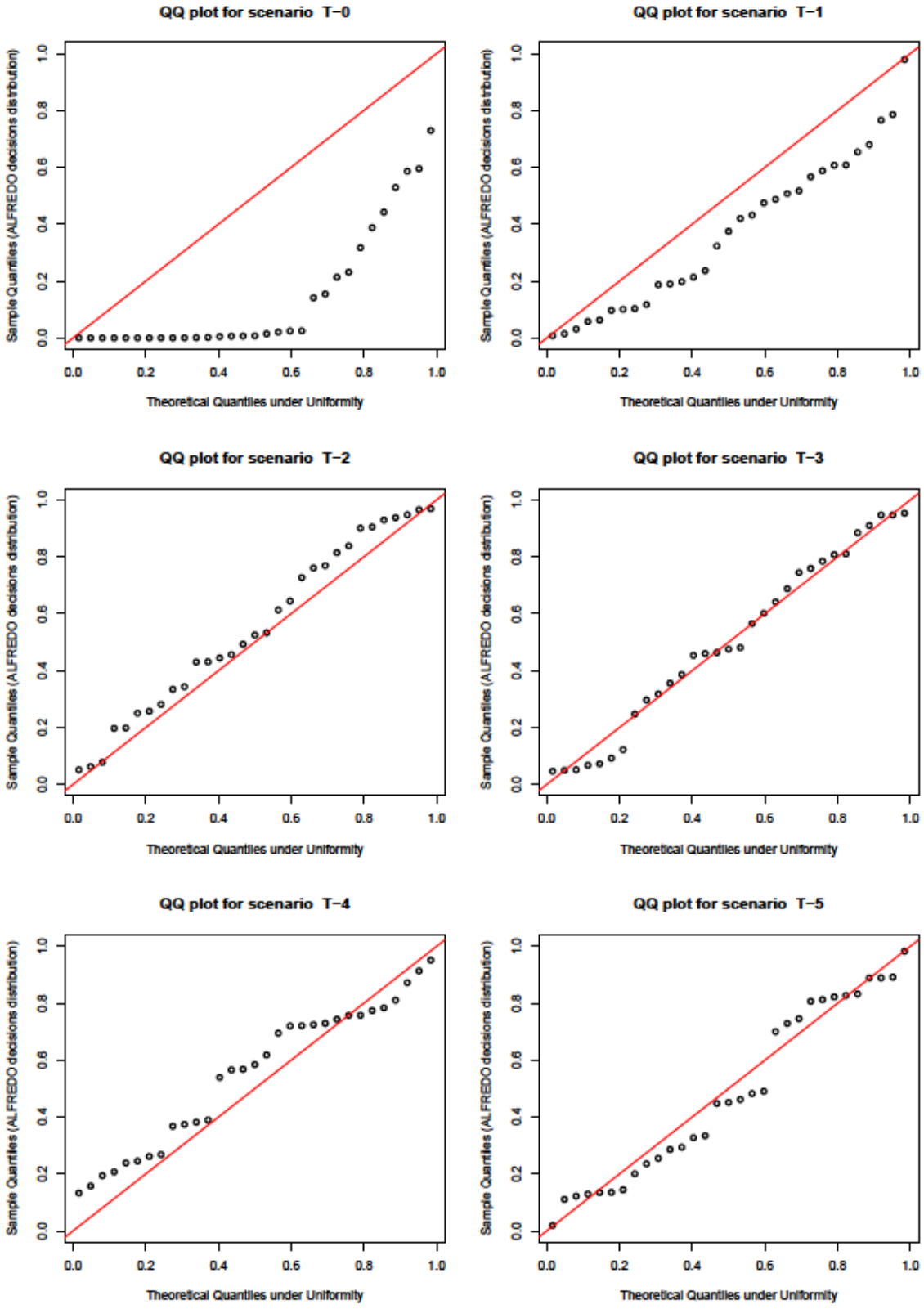



Figure 8.4: QQ-plot comparing distributions of ALFREDO back-testing results with uniform distribution. Note that for scenario T-0 and T-1 we can clearly reject the uniform distribution.

actually works and that the semantic knowledge base created from unstructured sources provides some insights into the next day companies' results. These observations allowed us to validate the remaining two hypotheses: H1 and H3.

At this point, the investigation process has been completed.

# 9 *Conclusions and future work*

If we compared this thesis to a sailing journey, we would be navigated by the lighthouses of the thesis' objectives. The investigation process (from Section 3.2), laid out like a sea map, would be our only guide towards the unknown. It would show us the directions to the nearest port of call with only a small promise of achievement. The journey would be very difficult and adventurous, but on the way, we would eventually make a few surprising discoveries. Those few breakthrough moments when we feel that our intuition was right and a hypothesis proved correct. And this is where we are now, in a safe haven of the *Conclusions* chapter.

This chapter summarises and concludes the research presented in this thesis. We recap main achievements and contributions of this work and explain how the hypotheses are validated. We juxtapose the thesis objectives with the main results that justify this work. After presenting the main outcomes we also describe how this work establishes a solid basis for the future lines of investigation.

## 9.1 Conclusions

When defining the objectives of this thesis, we stated three main problems: (i) the necessity of including new data sources in the process of financial analysis, (ii) the growing body of available unstructured data and (iii) the difficulty in unleashing the potential of unstructured data due to the lack of interoperability and computational form.

These problems were further contextualised in the domain of Business Intelligence and Decision Support Systems and based on that we devised decision-making process that utilises the unstructured data and domain ontologies. What is noteworthy, is that we decided to pursue this goal in the context of the Spanish market that implies focusing only on the textual sources in the Spanish language.

Creating an automated DSS require several preconditions to be thoroughly studied that later formed the main pillars of this work. Those are: (i) Natural Language Processing approaches to Information Extraction from unstructured data, (ii) Semantic Modelling for knowledge representation and (iii) text analytics for the financial decision-making process. We performed a state-of-the-art analysis on those three research areas where we identified the most relevant work, but also some of their shortcomings and open possibilities for improvement.

Based on that, we defined four research hypotheses to pursue and confirm or reject them in the course of this thesis. For this reason, we defined the investigation process that is a methodological underpinning for this research. Moreover, this investigation process is closely related to the main objectives, as stated in the inception of this thesis.

In Chapter 4.2 we analysed various sources of unstructured data for the financial decision-making process. We chose to use news and company disclosures. We performed the complete process of data acquisition that comprises the crawling of news articles, metadata extraction, test preprocessing and boilerplate removal. We evaluated the resulting corpus and indexed it for further processing. In the course of this work, we achieved the **Objective 1**, by creating a clean-text corpus of around

2.9 million documents.

Later, in Chapter 5 we perform two main tasks. First, we perform an analysis of the financial events of interests, based on existing literature and other sources (such as *hechos relevantes* – a similar to 8-K SEC filings that announce major events that stakeholder should know about). Then, based on that, we create a semantic model for representing financial knowledge in terms of well-defined concepts and relations. The semantic model is created with the idea to be later used for representing facts extracted from unstructured data. Here we address two objectives: **Objective 2**, by creating a list of financial events of high relevance for the financial decision-making and **Objective 3** by defining the semantic model for financial event representation. We also populate the dataset with some preliminary information, such as company list.

Chapter 6 is where the previous objectives prove their high relevance. We design the information extraction process, based on the novel technique using Convolutional Neural Network classifier and previous results (i.e. event definition and semantic model). The extracted events are later represented in a form of semantic dataset with the goal to be used as input to the decision model. This work fulfill the **Objective 4** and what is most important, it also validates two hypotheses: **H2** and **H4** (more on that later).

Having those results so far we were ready for the final goal: the design of a Decision Support System that would exploit all the previous results and prove its usefulness in the evaluation process. This is accomplished in Chapter 7 where the ALFREDO DSS is devised and Chapter 8 where this system is evaluated against the real-world data. By showing the positive results on the Spanish market, we met **Objective 5**. At the same time, we validated hypotheses: **H1** and **H3** (which we detail later).

## 9.2 *Hypotheses validation summary*

The investigation process devised in Section 3.2 and further work as explained in Section 9.1 lead us to the validation of our four research hypothesis:

*H1:* **Automated ontology-based information extraction of unstructured financial data leads to better decision-making process in the financial domain, and Business Intelligence in particular.**
We performed quantitative validation (see Section 8.4)of the Decision Support System (see Chapter 7) that uses unstructured data sources (see Chapter 4) and ontologies (see Chapter 5) in the context of the Spanish market. We validate this hypothesis by achieving better return than the market average, as represented by the IBEX35 index. This hypothesis is in a way similar to hypothesis H3, but is stronger as it implies the improvement of the actual decision-making. We show such improvement in the case of the "predictive" scenario $T-1$ (see Section 8.3) and we further confirm the soundness of the decision model (explained in Section 8.5).

*H2:* **Encoding semantic facts in the process of machine learning improves text classification:**

- **Knowledge-based systems help improving the process of extracting information from unstructured financial data,**
- **Semantic technologies improve accuracy in mining unstructured financial data.**

We improved the existing process of text classification with neural networks by adding semantic annotations in a form of ontology class embeddings (see Section 6.5). This approach lowered the error rate as compared to the previous variant without such annotations. This was possible as we already use semantic knowledge base in the decision-making process to perform ontology-based NER (see Sections 6.1 and Section 5), which satisfies the definition of a knowledge-based system.

We also show that the accuracy is higher when adding semantic annotations (see Section 6.6.4).

*H3:* **Semantic knowledge base created from unstructured sources improves analytical capabilities in the financial domain.**
The ALFREDO Decision Support System is using extracted semantic features (see Chapter 7) as a core idea of the structured decision support model. As shown in the quantitative evaluation in Section 8.4 thanks to this approach, the model can automatically learn from the past data in order to recognize and correctly classify the impact of real-world financial events 67% of time in the $T - 0$ scenario.

*H4:* **The use of deep learning techniques in information extraction can improve the process of text analysis in the financial domain as compared to the traditional machine learning methods.**
We performed a comparison of traditional techniques versus the novel deep learning approach against our corpora (See Section 6.6). The result of the evaluation (see Section 6.6.4) favours the deep learning approach, while using only named entity (semantic) annotations (see Section 6.5). Even not using any hand-engineered features gives a better results over the traditional approach.

## 9.3 *Thesis contributions*

The main outcome of this thesis is the ALFREDO Decision Support System that confirms the viability of using unstructured textual sources in the process of decision-making in finance. Also, in the course of this thesis, we produced other valuable contributions. The following list presents recaps the most important ones:

1. The news corpus of around 2.9 million documents in clean text from main Spanish news sites (see Appendix A for a complete source list). This corpus can be reused for other tasks that involve analysis of news and general text, and it has clean-text form that allows to use it without a need for laborious pre-processing. It covers the thematic scope

of general news: politics, economy, society, sports, with the emphasis on economy and politics.

2. Unique linguistic resources for financial event extraction (in Spanish). This can be used for training other classifiers and comparing results of different approach to information extraction in the financial domain.

3. Ontology for representing Spanish market data (companies, stocks, indices, financial events, roles, etc.). The semantic conceptual models that allows for semantic modelling of financial events, and potentially extends to other financial facts. It has been aligned with other well-known vocabularies and ontologies in order to provide further interoperability and to foster data publication, e.g. in the LOD cloud.

4. Semantic dataset with extracted financial events (populated semantic knowledge base). Most probably the first semantic dataset that covers the domain of financial events on the Spanish market. This dataset was successfully used to train a DSS model with a promising results. However the dataset can be used in many other ways.

5. Financial event classification models (trained Convolutional Neural Networks). A set of classifiers for extracting atomic facts from news texts. While the use case scenario for those resources is rather straightforward, it is also important to mention that whose can be used in other context, such as in medical domain, by previously train with domain-specific corpora.

6. DSS model for the financial-making process. The complete architecture that allows including unstructured data in the process of decision-making in the context of finance. This also comprises all stages of the data flow, starting from raw textual sources and ending on actionable information.

The obtained results of this thesis can be used for reproducing original results, starting new investigations, or further advance this work without having to start everything from scratch. These results can be combined

or exploited separately and bootstrap new investigation processes along these lines.

## *9.4 Future research*

The work accomplished within this thesis provides a solid ground for future research. There are multiple ways in how this research can be continued or extended into new grounds:

- From the point of view of unstructured sources we can look into other promising classes of data sources, either textual (such as blogs and other social media) or non-textual (videos, images and other formats). This could open a new possibility for including even more sources of data that can sometimes carry some kinds of information faster than other, e.g. textual sources. On the other hand, the inclusion of new sources could potentially trigger a new line of research that could focus on the reliability and veracity of unstructured sources in the decision-making process.
- The use of the semantic knowledge base lowers the barrier of data integration. This allows data augmentation or data blending with other datasets, possibly containing features that could improve the training of the current decision model. We could envisage the inclusion of data that can directly influence company results, such as e.g. raw material prices, but also more global ones, such as macro-economic indicators. The availability of such data could not only give more insights but also extend the predictive possibilities of the DSS.
- In the current research we were focused on the sentence-level information extraction. This obviously sets limits to the cognitive capabilities of the system, as we can not detect any fact that spans two sentences. We could however extend the financial facts classification to the paragraph or document-level. Such approach could further allow us to spot relevant knowledge in a broader context, not only in separate sentences. This way we could spot more fine-grained details about financial events and improve the classification in cases when a single

sentence can have a wrong interpretation when lacking a context in which it appears.

- Another interesting possibility would be to investigate the idea of end-to-end neural network models that could automatically learn relevant high-level features on their own, without the need of training information extraction classifiers, as is the case of the ALFREDO DSS. This could lower the threshold for creating new decision support systems almost from scratch, without the need for laborious feature engineering or availability of some specific linguistic resources. If the neural networks are so efficient in feature learning, this could be used to our advantage in order to construct more intelligent DSSs.
- The recent advancements in the field of deep learning and especially Recurrent Neural Networks could be brought to this area in a several ways: by further improving text classification and by advancing the decision support model by adding a temporal aspect of financial event chains, through the Long-Short Term Memory architecture. This could allow drawing conclusions from the temporal evolution of financial events, and look at some facts in conjunction with others, rather analysing each financial event separately.

We could also hypothesise about more brave and visionary ideas, even if they already feel very distant and risky or go across different domains. Such as extending the decision-support approach to analyse not only separate companies but the whole networks of interconnected financial entities. Analyse the systemic risk based on the currently unrolling events worldwide and, maybe, avoid the next financial catastrophe?

# *Bibliography*


Agichtein, E. and L. Gravano (2000). "Snowball: Extracting Relations from Large Plain-Text Collections". *Proceedings of the fifth ACM conference on Digital libraries - DL '00*, I(58), pp. 85–94. DOI: 10.1145/336597.336644.

Alić, I., J. Muntermann, and R. W. Gregory (2012). "State of the Art of Financial Decision Support Systems based on Problem, Requirement, Component and Evaluation Categories". In: *Proceeding of the 25th Conference Reliable and Trustworthy Structures, Process, Operations and Services for the Future*, pp. 280–293.

Alter, S. (1977). "A Taxonomy of Decision Support Systems". *Sloan Management Review*, 19(1), pp. 39–56.

Antweiler, W. and M. Frank (2006). "Do US stock markets typically overreact to corporate news stories?" *Working Paper*, pp. 1–22.

Auer, S., L. Bühmann, C. Dirschl, O. Erling, M. Hausenblas, R. Isele, J. Lehmann, M. Martin, P. N. Mendes, B. Van Nuffelen, C. Stadler, S. Tramp, and H. Williams (2012). "Managing the life-cycle of linked data with the LOD2 stack". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 7650 LNCS. PART 2, pp. 1–16. DOI: 10.1007/978-3-642-35173-0-1.

Auer, S., S. Dietzold, J. Lehmann, S. Hellmann, and D. Aumueller (2009). "Triplify: light-weight linked data publication from relational databases". In: *Proceedings of the 18th international conference on World wide web -*

*WWW '09*. New York, New York, USA: ACM Press, p. 621. DOI: 10.1145/1526709.1526793.

Baker, M. and J. Wurgler (2007). "Investor Sentiment in the Stock Market". *Journal of Economic Perspectives*, 21(2), pp. 129–151. DOI: 10.1257/jep.21.2.129.

Banko, M., M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni (2007). "Open Information Extraction from the Web". *Proceedings of IJCAI-07, the International Joint Conference on Artificial Intelligence*, pp. 2670–2676. DOI: 10.1145/1409360.1409378.

Bar-Haim, R., E. Dinur, R. Feldman, M. Fresko, and G. Goldstein (2011). "Identifying and following expert investors in stock microblogs". *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1310–1319.

Bennett, M. (2013). "The financial industry business ontology: Best practice for big data". *Journal of Banking Regulation*, 14(3-4), pp. 255–268. DOI: 10.1057/jbr.2013.13.

Bergeron, B. (2003). *Essentials of XBRL: Financial Reporting in the 21st Century*. Wiley, p. 240.

Bergstra, J. and Y. Bengio (2012). "Random Search for Hyper-Parameter Optimization". *Journal of Machine Learning Research*, 13, pp. 281–305.

Berners-Lee, T. (2010). *Linked data-design issues*.

Berners-Lee, T., J. Hendler, and O. Lassila (2001). "The Semantic Web". *Scientific American*, 284(5), pp. 34–43. DOI: 10.1038/scientificamerican0501-34.

Bird, S., E. Klein, and E. Loper (2009). *Natural Language Processing with Python*. Vol. 43, p. 479. DOI: 10.1097/00004770-200204000-00018.

Bizer, C., T. Heath, and T. Berners-Lee (2009). "Linked data-the story so far". *International journal on Semantic Web and Information Systems*, 5(3), pp. 1–22. DOI: 10.4018/jswis.2009081901.

Björkelund, A., B. Bohnet, L. Hafdell, and P. Nugues (2010). "A high-performance syntactic and semantic dependency parser". *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations (COLING 2010)*, (August), pp. 33–36.

Björne, J., J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, and T. Salakoski (2011). "Extracting contextualized complex biological events with rich graph-based feature sets". *Computational Intelligence*, 27(4), pp. 541–557. DOI: 10.1111/j.1467-8640.2011.00399.x.

Black, W. J., J. McNaught, A. Vasilakopoulos, K. Zervanou, B. Theodoulidis, and F. Rinald (2005). "CAFETIERE: Conceptual Annotations for Facts, Events, Terms, Individual Entities, and RElations". *Parmenides Technical Report TR-U4.3.1*.

Blazquez, M., M. Fernández, J. M. García-Pinar, and A. Gómez-Pérez (1998). "Building Ontologies at the Knowledge Level using the Ontology Design Environment". *11th International Workshop on Knowledge Acquisition Modeling and Management KAW98*, SHARE 4.1 –4.15.

Bollen, J., H. Mao, and X. Zeng (2011). "Twitter mood predicts the stock market". *Journal of Computational Science*, 2(1), pp. 1–8. DOI: 10.1016/j.jocs.2010.12.007.

Bontcheva, K., V. Tablan, D. Maynard, and H. Cunningham (2004). "Evolving GATE to meet new challenges in language engineering". *Natural Language Engineering*, 10(3-4), pp. 349–373. DOI: 10.1017/S1351324904003468.

Borst, W. N. (1997). *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. Vol. PhD, p. 243.

Brickley, D. and R. Guha (2008). *RDF Schema 1.1 - W3C Recommendation*. DOI: 10.1016/B978-0-12-373556-0.00006-X.

Brin, S. (1999). "Extracting Patterns and Relations from the World Wide Web". In: *The World Wide Web and Databases*. Vol. 53. 9, pp. 172–183. DOI: 10.1017/CBO9781107415324.004.

Brooks, F. P. J. (1995). *The Mythical Man-Month: Essays on Software Engineering, Anniversary Edition (2nd Edition)*. Addison-Wesley Professional;

Brown, G. W. and M. T. Cliff (2005). "Investor Sentiment and Asset Valuation". *The Journal of Business*, 78(2), pp. 405–440. DOI: 10.1086/427633.

Brown, S. J. and Warner, J. (1980). "Measuring security price perfomance". *Journal of Financial Economics*, 8, 8: 205–258. DOI: 10.1016/0304-405X(80)90002-1.

Brown, S. J. and Warner, J. (1985). "Using daily stock returns. The case of event studies". *Journal of Financial Economics*, 14(1), pp. 3–31. DOI: 10.1016/0304-405X(85)90042-X.

Budin, G., I. Soukup-Unterweger, and G. Sauberer (2010). "Terminology and Ontology Interoperability Model for Internal Financial Control Assessor Learning Environment". In: *Proceedings of the MONTIFIC Project at the Conference of The Current Financial Crisis and Competences to Address Problems on the European Market*, p. 57.

Chowdhuri, R., V. Y. Yoon, R. T. Redmond, and U. O. Etudo (2014). "Ontology based integration of XBRL filings for financial decision making". *Decision Support Systems*, 68, pp. 64–76. DOI: 10.1016/j.dss.2014.09.004.

Chung, W., H. Chen, and J. Nunamaker J.F. (2003). "Business intelligence explorer: a knowledge map framework for discovering business intelligence on the Web". In: *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on*, 10 pp.–. DOI: 10.1109/HICSS.2003.1173649.

Chung, W. (2014). "BizPro: Extracting and categorizing business intelligence factors from textual news articles". *International Journal of Information Management*, 34(2), pp. 272–284. DOI: 10.1016/j.ijinfomgt.2014.01.001.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa (2011). "Natural Language Processing (Almost) from Scratch". *The Journal of Machine Learning Research*, 12, pp. 2493–2537.

Connolly, D., F. V. Harmelen, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein (2001). "DAML+OIL (March 2001) Refer-

ence Description". *W3C Note 18 December 2001*, http://www.w3.org/TR/daml+oil–reference.

Cox, S. and C. Little (2012). *Time Ontology in OWL*.

Cunningham, H. (2002). "GATE, a general architecture for text engineering". *Computers and the Humanities*, 36(2), pp. 223–254. DOI: 10.1023/A:1014348124664.

Cyganiak, R., D. Wood, and M. Lanthaler (2014). *RDF 1.1 Concepts and Abstract Syntax*. Tech. rep. February 2014, pp. 1–22. DOI: 10.1007/s13398-014-0173-7.2.

Davies, J., D. Fensel, and F. Van Harmelen (2003). *Towards the Semantic Web: Ontology-Driven Knowledge Management*, xx, 288 p. DOI: 10.1017/S0269888905000305.

Da, Z., J. Engelberg, and P. Gao (2011). "In Search of Attention". *Journal of Finance*, 66(5), pp. 1461–1499. DOI: 10.1111/j.1540-6261.2011.01679.x.

Debreceny, R., S. Farewell, M. Piechocki, C. Felden, and A. Gräning (2010). "Does it add up? Early evidence on the data quality of XBRL filings to the SEC". *Journal of Accounting and Public Policy*, 29(3), pp. 296–306. DOI: 10.1016/j.jaccpubpol.2010.04.001.

Declerck, T., H. Krieger, S. Thomas, P. Buitelaar, T. Wunner, G. Maguet, J. McCrae, D. Spohr, and E. Montiel-Ponsoda (2010). "Ontology-based Multilingual Access to Financial Reports for Sharing Business Knowledge across Europe". *Internal Financial Control Assessment Applying Multilingual Ontology Framework*, (October), pp. 67–76.

Dhar, V. (2012). "Data Science and Prediction". *Communications of the ACM*, 56(12), pp. 64–73. DOI: 10.2139/ssrn.2086734.

Dublin Core Metadata Initiative (2012). *DCMI metadata terms specification*. DOI: 10.1108/10650750210418190.

Dzielinski, M., M. O. Rieger, and T. Talpsepp (2012). "Volatility asymmetry, news, and private investors". In: *The Handbook of News Analytics in*

*Finance*. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., pp. 255–270. DOI: 10.1002/9781118467411.ch11.

Eckart de Castilho, R. (2013). "Apache UIMA, Apache uimaFIT, and DKPro Core Tutorial". In: *3rd UIMA@GSCL Workshop, GSCL 2013, Darmstadt*.

Eckart de Castilho, R. and I. Gurevych (2014a). "A broad-coverage collection of portable NLP components for building shareable analysis pipelines". In: *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, pp. 1–11.

Eckart de Castilho, R. and I. Gurevych (2014b). "A broad-coverage collection of portable NLP components for building shareable analysis pipelines". In: *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pp. 1–11.

Essid, M., O. Boucelma, F.-M. Colonna, and Y. Lassoued (2004). "Query processing in a geographic mediation system". *Proceedings of the 12th annual ACM international workshop on Geographic information systems - GIS '04*, p. 101. DOI: 10.1145/1032222.1032239.

Ester, M., H. P. Kriegel, J. Sander, and X. Xu (1996). "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". *Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231. DOI: 10.1.1.71.1980.

Etzioni, O., M. Cafarella, D. Downey, A. M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates (2005). "Unsupervised named-entity extraction from the Web: An experimental study". *Artificial Intelligence*, 165(1), pp. 91–134. DOI: 10.1016/j.artint.2005.03.001.

Fani, H. and E. Bagheri (2015). "An Ontology for Describing Security Events". In: *The 27th International Conference on Software Engineering and Knowledge Engineering, {SEKE} 2015, Wyndham Pittsburgh University Center, Pittsburgh, PA, USA, July 6-8, 2015*, pp. 455–460. DOI: 10.18293/SEKE2015-101.

Fensel, D., I. Horrocks, F. Van Harmelen, S. Decker, M. Erdmann, M. Klein, F. V. Harmelen, S. Decker, M. Erdmann, and M. Klein (2000). "OIL in a nutshell". *Knowledge Engineering and Knowledge Management Methods, Models, and Tools*, 1937(1937), pp. 137–154. DOI: 10.1007/3-540-39967-4_1.

Ferguson, R. L. and C. H. Jones (1969). "A Computer Aided Decision System". *Management Science*, 15(10), DOI: 10.1287/mnsc.15.10.B550.

Fernández-López, M. (1999). "Overview Of Methodologies For Building Ontologies". *Proceedings of the IJCAI99 Workshop on Ontologies and ProblemSolving Methods Lessons Learned and Future Trends CEUR Publications*, 1999(2), pp. 1–13. DOI: 10.1.1.39.6002.

Fernández-López, M., A. Gómez-Pérez, and N. Juristo (1997). "METHONTOLOGY: From Ontological Art Towards Ontological Engineering". *AAAI-97 Spring Symposium Series*, SS-97-06, pp. 33–40. DOI: 10.1109/AXMEDIS.2007.19.

Ferrucci, D. and A. Lally (2004). "UIMA: an architectural approach to unstructured information processing in the corporate research environment". *Natural Language Engineering*, 10(3-4), pp. 327–348. DOI: 10.1017/S1351324904003523.

Finkel, J., T. Grenager, and C. Manning (2005). "Incorporating non-local information into information extraction systems by Gibbs sampling". *ACL '05 Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, (June), pp. 363–370. DOI: 10.3115/1219840.1219885.

Fox, M. S. (1992). "The TOVE Project: Towards a Common-sense Model of the Enterprise". In: *Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pp. 25–34. DOI: 10.1007/BFb0024952.

Fox, M. S., M. Grüninger, and M. Gruninger (1997). "On Ontologies and Enterprise Modelling". *Proc. Int'l Conf. Enterprise Integration Modeling Technology*, pp. 109–121. DOI: 10.1007/978-3-642-60889-6_22.

Gangemi, A., N. Guarino, C. Masolo, A. Oltramari, and L. Schneider (2002). "Sweetening ontologies with DOLCE". *Knowledge Engineering*

*and Knowledge Management: Ontologies and the Semantic Web, Lecture Notes in Computer Science, vol. 2473*, pp. 223–233. DOI: 10.1007/3-540-45810-7_18.

Gangemi, A. and P. Mika (2003). "Understanding the Semantic Web through Descriptions and Situations". *Proceedings of ODBASE03 Conference*, pp. 689–706.

Gantz, J., D. Reinsel, and B. D. Shadows (2012). "The Digital Universe in 2020". *IDC iView "Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East"*, 2007(December 2012), pp. 1–16.

Gao, Y. B. and G. Zhao (2005). "Knowledge-based information extraction: A case study of recognizing emails of Nigerian frauds". *Natural Language Processing and Information Systems, Proceedings*, 3513, pp. 161–172.

García, R. and R. Gil (2010). "Linking XBRL financial data". In: *Linking Enterprise Data*, pp. 103–125. DOI: 10.1007/978-1-4419-7665-9_6.

Geva, T. and J. Zahavi (2014). "Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news". *Decision Support Systems*, 57(1), pp. 212–223. DOI: 10.1016/j.dss.2013.09.013.

Gibson, C. H. (2012). *Financial Reporting and Analysis: Using Financial Accounting Information*. 13th editi. South-Western Cengage Learning, 688 pages.

Gilbert, E. and K. Karahalios (2010). "Widespread Worry and the Stock Market". *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pp. 58–65. DOI: 10.1.1.220.9231.

Goldberg, Y. and O. Levy (2014). "word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method". *arXiv preprint arXiv:1402.3722*, (2), pp. 1–5.

Gómez-Berbís, J. M., F. García-Sánchez, R. Valencia-García, I. Toma, and C. G. Moreno (2009). "SONAR: A Semantically Empowered Financial

Search Engine". In: Springer Berlin Heidelberg, pp. 405–414. DOI: 10.1007/978-3-642-02264-7_42.

Gómez-Pérez, A., M. Fernández-López, and O. Corcho (2004). *Ontological Engeenering*. Advanced Information and Knowledge Processing. London: Springer-Verlag, p. 403. DOI: 10.1007/b97353.

Gómez-Pérez, A. and M. C. Suárez-Figueroa (2009). "Scenarios for building ontology networks within the NeOn methodology". In: *Proceedings of the fifth international conference on Knowledge capture - K-CAP '09*. New York, New York, USA: ACM Press, p. 183. DOI: 10.1145/1597735.1597773.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. MIT Press.

Goodrum, A. A. (2000). "Image information retrieval: An overview of current research". *Informing Science*, 3(2), pp. 63–67.

Goonatilake, R. and S. Herath (2007). "The Volatility of the Stock Market and News". *International Research Journal of Finance and Economics*, (Issue 11 (2007)), pp. 53–65.

Goth, G. (2016). "Deep or Shallow, NLP is Breaking out". *Communications of the ACM*, 59(3), pp. 13–16. DOI: 10.1145/2874915.

Government of Spain (1988). "Ley 24/1988, de 28 de Julio, del Mercado de Valores." *Boletín Oficial de Estado*, 310, pp. 105198–105294.

Grčar, M., P. Kralj, D. Velizar, and A. Klein (2012). "Ontology reuse and evolution". *Project FIRST deliverable D3.2*.

Grishman, R. and B. Sundheim (1996). "Message Understanding Conference-6: A Brief History". *Proceedings of the 16th conference on Computational linguistics*, 1, pp. 466–471. DOI: 10.3115/992628.992709.

Groth, S. S. and J. Muntermann (2011). "An intraday market risk management approach based on textual analysis". *Decision Support Systems*, 50(4), pp. 680–691. DOI: 10.1016/j.dss.2010.08.019.

Gruber, T. R. (1993). "A translation approach to portable ontology specifications". *Knowledge Acquisition*, 5(2), pp. 199–220. DOI: 10.1.1.101.7493.

Gruber, T. R. (1995a). "Toward principles for the design of ontologies used for knowledge sharing". *International Journal of Human-Computer Studies*, 43(5-6), pp. 907–928. DOI: citeulike-article-id:230211.

Gruber, T. R. (1995b). "Toward principles for the design of ontologies used for knowledge sharing". *International Journal of Human-Computer Studies*, 43(5-6), pp. 907–928. DOI: citeulike-article-id:230211.

Hackathorn, R. D. and P. G. W. Keen (1981). "Organizational Strategies for Personal Computing in Decision Support Systems". *MISQ*, 5(3), p. 21. DOI: 10.2307/249288.

Hausenblas, M. (2010). "The Linked Open Data Around-the-Clock project Fact Sheet".

Hausenblas, M., B. Villazón-Terrazas, and B. Hyland (2011). "GLD Life Cycle." *W3C Government Linked Data Group. W3C*.

Heather, T. and C. Bizer (2011). *Linked Data - Evolving the Web into Global Data Space*. Vol. 5, pp. 1–6. DOI: 10.2200/S00334ED1V01Y201102WBE001.

Hendler, J. (2001). "Agents and the semantic web". *IEEE Intelligent Systems and Their Applications*, 16(2), pp. 30–37. DOI: 10.1109/5254.920597.

Hepp, M. (2007). "Possible ontologies: How reality constrains the development of relevant ontologies". *IEEE Internet Computing*, 11(1), pp. 90–96. DOI: 10.1109/MIC.2007.20.

Hogenboom, A., F. Hogenboom, F. Frasincar, K. Schouten, and O. Van Der Meer (2013). "Semantics-based information extraction for detecting economic events". *Multimedia Tools and Applications*, 64(1), pp. 27–52. DOI: 10.1007/s11042-012-1122-0.

Hogenboom, A., F. Hogenboom, U. Kaymak, P. Wouters, and F. De Jong (2010). "Mining economic sentiment using argumentation structures". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in*

*Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 6413 LNCS, pp. 200–209. DOI: 10.1007/978-3-642-16385-2_25.

Hogenboom, F., F. Frasincar, U. Kaymak, and F. De Jong (2011). "An overview of event extraction from text". In: *CEUR Workshop Proceedings*. Vol. 779, pp. 48–57.

Holsapple, C. W. and A. B. Whinston (1996). *Decision support systems: a knowledge-based approach*. West Pub. Co.

Horrocks, I. (2002). "DAML + OIL : a Description Logic for the Semantic Web". *Bull. of the IEEE Computer Society Technical Committee on Data Engineering*, 25, pp. 1–7.

Horst, H. (2005). "Combining RDF and part of OWL with rules: Semantics, decidability, complexity". *The Semantic Web–ISWC 2005*, pp. 668–684.

Hotho, A., A. Nürnberger, and G. Paaß (2005). "A Brief Survey of Text Mining". *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 20, pp. 19–62. DOI: 10.1111/j.1365-2621.1978.tb09773.x.

Hu, B., Z. Lu, H. Li, and Q. Chen (2014). "Convolutional Neural Network Architectures for Matching Natural Language Sentences". *Advances in Neural Information Processing Systems 27*, pp. 2042–2050.

Java, A., T. Finin, and S. Nirenburg (2006). "SemNews: A Semantic News Framework". In: *In Proceedings of the Twenty-First National Conference on Artificial Intelligence*.

Jurafsky, D. S. and J. H. Martin (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. DOI: 10.1162/089120100750105975.

Kahneman, D. and A. Tversky (1979). "Prospect Theory: An Analysis of Decision under Risk". *Econometrica*, 47(2), pp. 263–292. DOI: 10.2307/1914185.

Kakkonen, T. and T. Mufti (2011). "Developing and applying a company, product and business event ontology for text mining". *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies - i-KNOW '11*, p. 1. DOI: 10.1145/2024288.2024318.

Kalchbrenner, N., E. Grefenstette, and P. Blunsom (2014). "A Convolutional Neural Network for Modelling Sentences". *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, {ACL} 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pp. 655–665.

Kerremans, K., Y. Tang, R. Temmerman, and G. Zhao (2005). "Towards Ontology-based E-mail Fraud Detection". In: *2005 Purtuguese Conference on Artificial Intelligence*. IEEE, pp. 106–111. DOI: 10.1109/EPIA.2005.341275.

Kim, Y. (2014). "Convolutional Neural Networks for Sentence Classification". *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 1746–1751.

Kim, Y., Y. Jernite, D. Sontag, and A. M. Rush (2016). "Character-Aware Neural Language Models". *Thirtieth AAAI Conference on Artificial Intelligence*.

Klein, M. and L. Methlie (2009). "Knowledge-Based Decision Support Systems With Applications in Business: A Decision Support Approach". *CiteULike Group Information Networks and Knowledge Management library*.

Kocianski, S. (2016). *The Robo-Advising Report: Market forecasts, key growth drivers, and how automated asset management will change the advisory industry*. Tech. rep. BI Intelligence.

Kogan, S., D. Levin, B. R. Routledge, J. S. Sagi, and N. a. Smith (2009). "Predicting Risk from Financial Reports with Regression". *Proceedings of ACL Human Language Technologies*, (June), pp. 272–280. DOI: 10.3115/1620754.1620794.

Kohlschütter, C., P. Fankhauser, and W. Nejdl (2010). "Boilerplate Detection using Shallow Text Features". *Text*, pp. 441–450. DOI: 10.1145/1718487.1718542.

Laender, A. H. F., B. a. Ribeiro-Neto, A. S. da Silva, and J. S. Teixeira (2002). "A brief survey of web data extraction tools". *ACM SIGMOD Record*, 31(2), p. 84. DOI: 10.1145/565117.565137.

Lara, R., I. Cantador, and P. Castells (2006). "XBRL taxonomies and OWL ontologies for investment funds". In: *25th International Conference on Conceptual Modelling*, pp. 6–9. DOI: 10.1007/11908883_33.

LeCun, Y., Y. Bengio, and G. Hinton (2015). "Deep learning". *Nature*, 521(7553), pp. 436–444. DOI: 10.1038/nature14539.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner (1998). "Gradient Based Learning Applied to Document Recognition". *Proceedings of the IEEE*, 86(11), pp. 2278–2324. DOI: 10.1109/5.726791.

Lee, H., M. Surdeanu, B. Maccartney, and D. Jurafsky (2014). "On the Importance of Text Analysis for Stock Price Prediction". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by N. C. ( Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis. Reykjavik, Iceland: European Language Resources Association (ELRA).

Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, and C. Bizer (2015). "DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia". *Semantic Web*, 6(2), pp. 167–195. DOI: 10.3233/SW-140134.

Leibold, C., H. U. Krieger, and M. Spies (2010). "Ontology-Based Modelling and Reasoning in Operational Risks". In: *Operational Risk Management: A Practical Approach to Intelligent Data Analysis*, pp. 41–59. DOI: 10.1002/9780470972571.ch3.

Leinweber, D. and J. Sisk (2011). "Event-Driven Trading and the "New News"". *The Journal of Portfolio Management*, 38(1), pp. 110–124. DOI: 10.3905/jpm.2011.38.1.110.

Leinweber, D. and J. Sisk (2012). *The handbook of news analytics in finance*, pp. 147–172. DOI: 10.1002/9781118467411.ch6.

Leinweber, D. (2009). *Nerds on Wall Street. Math, Machines and Wired Markets*. Vol. 1994. 9, pp. 181–202. DOI: 10.2307/1317604.

Lenat, D. B. and R. V. Guha (1990). *Building large knowledge-based systems: representation and inference in the Cyc project*. Addison-Wesley Pub. Co, p. 372.

Levy, O. and Y. Goldberg (2014). "Linguistic regularities in sparse and explicit word representations". *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pp. 171–180.

Lösch, U. and N. Nikitina (2009). "The newsEvents Ontology: An Ontology for Describing Business Events". In: *Proceedings of the 2009 International Conference on Ontology Patterns - Volume 516*. WOP'09. Aachen, Germany, Germany: CEUR-WS.org, pp. 187–193.

Luhn, H. (1958). "A Business Intelligence System". *IBM Journal of Research and Development*, 2(4), pp. 314–319. DOI: 10.1147/rd.24.0314.

Maaten, L. van der and G. E. Hinton (2008). "Visualizing high-dimensional data using t-SNE". *Journal of Machine Learning Research*, 9, pp. 2579–2605. DOI: 10.1007/s10479-011-0841-3.

MacKinlay, A. (1997). "Event studies in economics and finance". *Journal of economic literature*, 35(1), pp. 13–39.

Magnus Knuth Jens Lehmann, D. K. T. S. and H. Sack (2015). "The DBpedia Events Dataset". In: *Proceedings of the ISWC 2015 Posters & Demonstrations Track*. Ed. by M. D. Serena Villata Jeff Z. Pan. 1486. Bethlehem, PA, USA: CEUR-WS.org.

Manning, C. D., J. Bauer, J. Finkel, S. J. Bethard, M. Surdeanu, and D. McClosky (2014). "The Stanford CoreNLP Natural Language Processing Toolkit". *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60.

Matuszek, C., J. Cabral, M. Witbrock, and J. Deoliveira (2006). "An introduction to the syntax and content of Cyc". *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, 3864(1447), pp. 44–49.

McKinsey & Company (2011). "Big data: The next frontier for innovation, competition, and productivity". *McKinsey Global Institute*, (June), p. 156. DOI: 10.1080/01443610903114527.

Micu, A., L. Mast, V. Milea, F. Frasincar, and U. Kaymak (2009). "Financial News Analysis Using a Semantic Web Approach". *Semantic Knowledge Management: An Ontology-Based Framework*, pp. 311–328. DOI: 10.4018/978-1-60566-034-9.

Mika, P. (2005). "Social networks and the semantic web". In: *Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on*. IEEE, pp. 285–291.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). "Distributed Representations of Words and Phrases and their Compositionality". *Nips*, pp. 1–9. DOI: 10.1162/jmlr.2003.3.4-5.951.

Mikolov, T., G. Corrado, K. Chen, and J. Dean (2013). "Efficient Estimation of Word Representations in Vector Space". *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pp. 1–12. DOI: 10.1162/153244303322533223.

Mikolov, T., W.-t. Yih, and G. Zweig (2013). "Linguistic regularities in continuous space word representations". *Proceedings of NAACL-HLT*, (June), pp. 746–751.

Montes, M. M., J. L. Bas, S. Bellido, O. Corcho, S. Losada, R. Benjamins, and J. Contreras (2005). *Financial Ontology, Project DIP, Deliverable D10.3*. Tech. rep.

Motik, B., B. C. Grau, I. Horrocks, Z. Wu, A. Fokoue, and C. Lutz (2012). *OWL 2 Web Ontology Language Profiles (Second Edition)*.

Nadeau, D. (2007). "A survey of named entity recognition and classification". *Linguisticae Investigationes*, (30), pp. 3–26. DOI: 10.1075/li.30.1.03nad.

Nadkarni, A. and D. Vesset (2015). "Worldwide Big Data Technology and Services 2015 – 2019 Forecast". *Market Forecast, IDC Research, Inc.* (March 2012), p. 34.

Nann, S., J. Krauss, and D. Schoder (2013). "Predictive Analytics On Public Data - The Case Of Stock Markets". In: *Proceedings of the 21st European Conference on Information Systems*, pp. 1–12.

Nassirtoussi, A. K., S. Aghabozorgi, T. Ying Wah, and D. Ngo Chek Ling (2014). *Text mining for market prediction: A systematic review*. DOI: 10.1016/j.eswa.2014.06.009.

Nguyen, T. H. and R. Grishman (2015). "Relation Extraction: Perspective from Convolutional Neural Networks". *Workshop on Vector Modeling for NLP*, pp. 39–48.

Nirenburg, S. and V. Raskin (2001). "Ontological semantics, formal ontology, and ambiguity". *Formal Ontology in Information Systems. IOS Press*, pp. 151–161. DOI: 10.1145/505168.505183.

Nofer, M. and O. Hinz (2015). "Using Twitter to Predict the Stock Market: Where is the Mood Effect?" *Business and Information Systems Engineering*, 57(4), pp. 229–242. DOI: 10.1007/s12599-015-0390-4.

Nuij, W., V. Milea, F. Hogenboom, F. Frasincar, and U. Kaymak (2014). "An automated framework for incorporating news into stock trading strategies". *IEEE Transactions on Knowledge and Data Engineering*, 26(4), pp. 823–835. DOI: 10.1109/TKDE.2013.133.

Ogren, P. and S. Bethard (2009). "Building Test Suites for {UIMA} Components". In: *Proceedings of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing (SETQA-NLP 2009)*. Boulder, Colorado: Association for Computational Linguistics, pp. 1–4.

O'Riain, S., E. Curry, and A. Harth (2012). "XBRL and open data for global financial ecosystems: A linked data approach". *International Journal of Accounting Information Systems*, 13(2), pp. 141–162. DOI: 10.1016/j.accinf.2012.02.002.

O'Riain, S., A. Harth, and E. Curry (2012). "Linked Data Driven Information Systems as an Enabler for Integrating Financial Data". In: *Information Systems for Global Financial Markets: Emerging Developments and Effects*, pp. 239–270. DOI: 10.4018/978-1-61350-162-7.ch010.

Pak, A. and P. Paroubek (2010). "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In: *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).

Pang, B. and L. Lee (2008). "Opinion Mining and Sentiment Analysis". *Foundations and Trends in Information Retrieval*, 2(1–2), pp. 1–135. DOI: 10.1561/1500000001.

Pasternack, J. and D. Roth (2009). "Extracting article text from the web with maximum subsequence segmentation". *Proceedings of the 18th international conference on World wide web - WWW '09*, p. 971. DOI: 10.1145/1526709.1526840.

Pennington, J., R. Socher, and C. D. Manning (2014). "GloVe: Global Vectors for Word Representation". *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

Perez, A. G. and V. R. Benjamins (1999). "Overview of Knowledge Sharing and Reuse Components : Ontologies and Problem-Solving Methods". *IJCAI-99 workshop on Ontologies and Problem-Solving Method (KRR5)*, pp. 1–15.

Pfleeger, S. L. (2009). *Software Engineering: Theory and Practice (4th Edition)*. Pearson; 4 edition (February 27, 2009), 792 pages.

Pinsker, R. and S. Li (2008). "Costs and benefits of XBRL adoption". *Communications of the ACM*, 51(3), pp. 47–50. DOI: 10.1145/1325555.1325565.

Pinto, H. S., S. Staab, and C. Tempich (2004). "DILIGENT : Towards a fine-grained methodology for DIstributed , Loosely-controlled and evolvInG Engineering of oNTologies". *16Th European Conference on Artificial Intelligence - Ecai*, pp. 393–397.

Pomikálek, J. (2011). "Removing boilerplate and duplicate content from web corpora". PhD thesis.

Power, D. J. (2007). "A Brief History of Decision Support Systems". *DSS-Resources.COM*, (March), p. 1.

Power, D. J. (2001a). "Supporting decision-makers: An expanded framework". *Proceedings of Informing Science and IT Education, 2001*, 1(June), pp. 1901–1915. DOI: 10.1109/JSAC.2006.877218.

Power, D. J. (2001b). "Supporting decision-makers: An expanded framework". *Proceedings of Informing Science and IT Education, 2001*, 1(June), pp. 1901–1915. DOI: 10.1109/JSAC.2006.877218.

Power, D. J., R. Sharda, and F. Burstein (2015). *Decision Support Systems*, pp. 1–4. DOI: 10.1007/978-90-481-9045-4.

Prud'hommeaux, E. and A. Seaborne (2008). "SPARQL Query Language for RDF". *W3C Recommendation*, 2009(January), pp. 1–106. DOI: citeulike-article-id:2620569.

Radzimski, M. and J. L. Sánchez-Cervantes (2012). "FLORA - Publishing Unstructured Financial Information in the Linked Open Data Cloud". *Joint Proceedings of the 1st International Workshop on Finance and Economics on the Semantic Web (FEOSW 2012), 5th International Workshop on REsource Discovery (RED 2012) and 7th International Workshop on Semantic Business Process Management (SBPM 2012) i*, pp. 31–37.

Radzimski, M., J. L. Sánchez-Cervantes, J. L. L. Cuadrado, and Á. García-Crespo (2014). "Predicting stocks returns correlations based on un-

structured data sources". In: *CEUR Workshop Proceedings*. Vol. 1240, pp. 87–96.

Radzimski, M., J. L. Sanchez-Cervantes, A. Garcia-Crespo, and I. Temiño-Aguirre (2014). "Intelligent Architecture for Comparative Analysis of Public Companies Using Semantics and XBRL Data". *International Journal of Software Engineering and Knowledge Engineering*, 24(05), pp. 801–823. DOI: 10.1142/S0218194014500314.

Ranco, G., D. Aleksovski, G. Caldarelli, M. Grčar, and I. Mozeti?? (2015). "The effects of twitter sentiment on stock price returns". *PLoS ONE*, 10(9). DOI: 10.1371/journal.pone.0138441.

Raymond, R. C. (1966). "Use of the Time-Sharing Computer in Business Planning and Budgeting". *Management Science*, 12(8), DOI: 10.1287/mnsc.12.8.B363.

Rizzo, G. (2011). "NERD : A Framework for Evaluating Named Entity Recognition Tools in the Web of Data". In: *Proceedings of the 11th International Semantic Web Conference ISWC2011*, pp. 1–4.

Rong, X. (2014). "word2vec Parameter Learning Explained". *arXiv:1411.2738*, pp. 1–19.

Roth, D. and W.-T. Yih (2004). "A linear programming formulation for global inference in natural language tasks". *Proceedings of the 8th Conference on Computational Natural Language Learning*, pp. 1–8.

Ruiz-Martínez, J. M., R. Valencia-García, and F. García-Sánchez (2012). "Semantic-based sentiment analysis in financial news". In: *CEUR Workshop Proceedings*. Vol. 862, pp. 38–51.

Russom, P. (2007). "BI Search and Text Analytics". *TDWI Best Practices Report*.

Ryan, P. and R. J. Taffler (2004). "Are Economically Significant Stock Returns and Trading Volumes Driven by Firm-specific News Releases?" *Journal of Business Finance & Accounting*, 31(1-2), pp. 49–82. DOI: 10.1111/j.0306-686X.2004.0002.x.

Saggion, H., A. Funk, D. Maynard, and K. Bontcheva (2007). "Ontology-based Information Extraction for Business Intelligence". *The Semantic Web*, 4825, pp. 843–856. DOI: 10.1007/978-3-540-76298-0_61.

Sahoo, S. S., W. Halb, S. Hellmann, K. Idehen, T. Thibodeau Jr, S. Auer, J. Sequeda, and A. Ezzat (2009). "A survey of Current approaches for mapping of relational databases to RDF". *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(1), p. 15. DOI: 10.1016/j.websem.2007.11.011.

Sánchez-Cervantes, J. L. (2015). "Linked Data para la Generación de Conocimiento Financiero a partir de la Extracción de Información Semiestructurada". PhD thesis. Universidad Carlos III de Madrid.

Sánchez-Rada, J. F. and C. A. Iglesias (2013). "Onyx: Describing emotions on the web of data". In: *CEUR Workshop Proceedings*. Vol. 1096, pp. 71–82.

Sánchez-Rada, J. F., M. Torres, C. A. Iglesias, R. Maestre, and E. Peinado (2014). "A Linked Data Approach to Sentiment and Emotion Analysis of Twitter in the Financial Domain". In: *Joint Proceedings of the Second International Workshop on Semantic Web Enterprise Adoption and Best Practice and Second International Workshop on Finance and Economics on the Semantic Web Co-located with 11th European Semantic Web Conference, WaSABi-FEOSW*. Vol. 1240. CEUR Workshop Proceedings.

Sánchez-Rada, J. F., G. Vulcu, C. A. Iglesias, and P. Buitelaar (2014). "EUROSENTIMENT: Linked data sentiment analysis". In: *CEUR Workshop Proceedings*. Vol. 1272, pp. 145–148.

Santos, C. N. dos, B. Xiang, and B. Zhou (2015). *Classifying Relations by Ranking with Convolutional Neural Networks*.

Schreiber, G. and Y. Raimond (2014). *RDF 1.1 Primer*. Tech. rep.

Schumaker, R. P. and H. Chen (2009). "Textual analysis of stock market prediction using breaking financial news". *ACM Transactions on Information Systems*, 27(2), pp. 1–19. DOI: 10.1145/1462198.1462204.

Sebastiani, F. (2002). "Machine learning in automated text categorization". *ACM Computing Surveys*, 34(1), pp. 1–47. DOI: 10.1145/505282.505283.

Segers, R., P. Vossen, M. Rospocher, L. Serafini, E. Laparra, and G. Rigau (2015). "ESO: A Frame based Ontology for Events and Implied Situations". In: *Proceedings of MAPLEX 2015*. Yamagata, Japan.

Shadbolt, N., W. Hall, and T. Berners-Lee (2006). "The semantic web revisited". *IEEE Intelligent Systems*, 21(3), pp. 96–101. DOI: 10.1109/MIS.2006.62.

Sharpe, W. F. (1994). "The Sharpe Ratio". *The journal of Portfolio Management*, 21(1), pp. 49–58. DOI: 10.3905/jpm.1994.409501.

Soru, T. and A.-C. N. Ngomo (2014). "A comparison of supervised learning classifiers for link discovery". In: *Proceedings of the 10th International Conference on Semantic Systems, {SEMANTICS} 2014, Leipzig, Germany, September 4-5, 2014*. Ed. by H. Sack, A. Filipowska, J. Lehmann, and S. Hellmann. ACM, pp. 41–44. DOI: 10.1145/2660517.2660532.

Spies, M. (2010). "An ontology modelling perspective on business reporting". *Information Systems*, 35(4), pp. 404–416. DOI: 10.1016/j.is.2008.12.003.

Sporny, M., G. Kellogg, and M. Lanthaler (2013). *JSON-LD 1.0 -A JSON-based Serialization for Linked Data.*

Sprenger, T. O., P. G. Sandner, A. Tumasjan, and I. M. Welpe (2014). "News or noise? Using twitter to identify and understand company-specific news flow". *Journal of Business Finance and Accounting*, 41(7-8), pp. 791–830. DOI: 10.1111/jbfa.12086.

Sprenger, T. O., A. Tumasjan, P. G. Sandner, and I. M. Welpe (2014). "Tweets and trades: The information content of stock microblogs". *European Financial Management*, 20(5), pp. 926–957. DOI: 10.1111/j.1468-036X.2013.12007.x.

Sprenger, T. O. and I. M. Welpe (2011). "News or Noise? The Stock Market Reaction to Different Types of Company-Specific News Events".

*Technische Universität München*, (January), p. 65. DOI: 10.2139/ssrn.1734632.

Staab, S., R. Studer, H. P. Schnurr, and Y. Sure (2001). "Knowledge processes and ontologies". *IEEE Intelligent Systems and Their Applications*, 16(1), pp. 26–34. DOI: 10.1109/5254.912382.

Stenetorp, P., S. Pyysalo, G. Topi, T. Ohta, S. Ananiadou, and J. Tsujii (2012). "BRAT : a Web-based Tool for NLP-Assisted Text Annotation". *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12)*, (Figure 1), pp. 102–107.

Storkenmaier, A., M. Müller, and C. Weinhardt (2010). *A Software Framework for a News Event Driven Simulation of Algorithmic Trading Strategies*. Vol. 11, pp. 45–53. DOI: 978-3-941875-31-9.

Studer, R., V. Benjamins, and D. Fensel (1998). "Knowledge engineering: Principles and methods". *Data & Knowledge Engineering*, 25(1-2), pp. 161–197. DOI: 10.1016/S0169-023X(97)00056-6.

Suárez-Figueroa, M. C., A. Gómez-Pérez, and M. Fernández-López (2012). "The NeOn Methodology for Ontology Engineering". In: *Ontology Engineering in a Networked World*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 9–34. DOI: 10.1007/978-3-642-24794-1_2.

Sundermeyer, M., H. Ney, and R. Schluter (2015). "From Feedforward to Recurrent LSTM Neural Networks for Language Modeling". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3), pp. 517–529. DOI: 10.1109/TASLP.2015.2400218.

Surdeanu, M., D. McClosky, M. Smith, A. Gusev, and C. Manning (2011). "Customizing an Information Extraction System to a New Domain". *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics, RELMS '11*, (Relms), pp. 2–10.

The Dublin Core Metadata Initiative (2010). *DCMI Metadata Terms*. Tech. rep., pp. 1–45.

Toribio, R., P. Martínez, and C. de Pablo-Sánchez (2010). "Evaluación de la Extracción de Entidades Nombradas de OpenCalais en castellano". *Procesamiento del Lenguaje Natural*, 45, pp. 287–290.

Treleaven, P., M. Galas, and V. Lalchand (2013). "Algorithmic trading review". *Communications of the ACM*, 56(11), pp. 76–85. DOI: 10.1145/2500117.

Turner, V., J. F. Gantz, D. Reinsel, and S. Minton (2014). "The digital universe of opportunities: Rich data and the increasing value of the internet of things". *IDC Analyze the Future*.

Vanderlinden, E. (2011). *Fadyart Finance Ontology v4.00*.

Van Nuffelen, B., V. Janev, M. Martin, V. Mijovic, and S. Tramp (2014). "Supporting the Linked Data Life Cycle Using an Integrated Tool Stack". In: *Lecture Notes in Computer Science*. Springer International Publishing, pp. 108–129. DOI: 10.1007/978-3-319-09846-3_6.

W3C (2012). "OWL 2 Web Ontology Language Document Overview ( Second Edition )". *Ontology, Web Document, Language Edition, Second Recommendation, Latest Group, O W L Working Reserved, All Rights*, (December), pp. 1–7.

Wei Hu, Jianfeng Chen, and Yuzhong Qu (2011). *self-training approach to leveraging the semantics-based and similarity-based ways for addressing the problem of object coreference resolution on the Semantic Web*. Hyderabad, India.

Westerski, A., C. A. Iglesias, and F. T. Rico (2011). "Linked opinions: Describing sentiments on the structured web of data". In: *CEUR Workshop Proceedings*. Vol. 830.

West, M. (1999). "Developing High Quality Data Models". *The European Process Industries STEP Technical Liaison Executive (EPISTLE)*, p. 62.

Wilton, P., J. Tarling, and J. McGinnis (2013). *Storyline Ontology*.

Wu, J. (2017). *Introduction to Convolutional Neural Networks*. Tech. rep., p. 31.

Wysocki, P. (1998). *Cheap talk on the web: The determinants of postings on stock message boards*. DOI: 10.2139/ssrn.160170.

Xie, B., R. J. Passonneau, L. Wu, and G. G. Creamer (2013). *Semantic Frames to Predict Stock Price Movement*.

Zaki, M. and B. Theodoulidis (2013). "An Ontology for Monitoring and Surveillance Systems in Financial Markets". *SSRN Electronic Journal*. DOI: 10.2139/ssrn.2266679.

Zeng, D., K. Liu, S. Lai, G. Zhou, and J. Zhao (2014a). "Relation Classification via Convolutional Deep Neural Network". In: *The 25th International Conference on Computational Linguistics (COLING 2014)*. 2011, pp. 2335–2344. DOI: http://aclweb.org/anthology/C/C14/C14-1220.pdf.

Zeng, D., K. Liu, S. Lai, G. Zhou, and J. Zhao (2014b). "Relation Classification via Convolutional Deep Neural Network". *Coling*, (2011), pp. 2335–2344. DOI: http://aclweb.org/anthology/C/C14/C14-1220.pdf.

Zhao, G., J. Kingston, K. Kerremans, F. Coppens, R. Verlinden, R. Temmerman, and R. Meersman (2004). "Engineering an Ontology of Financial Securities Fraud". *On the Move to Meaningful Internet Systems 2004: OTM 2004 Workshops*, 3292, pp. 605–620.

# Appendices

# A List of financial news sources

Crawling for financial news in Spanish in the period of 1 January 2014 – 15 May 2016 from the following sources:

1. General news sites

   - ABC (http://abc.es)
   - Agencia EFE (http://www.efe.com)
   - Cinco Días (http://cincodias.com/)
   - Diario Abierto (http://www.diarioabierto.es/)
   - Diario Público (http://www.publico.es/)
   - El Comercio (http://www.elcomercio.es/)
   - El Confidential (http://www.elconfidencial.com/)
   - eldiario.es (http://www.eldiario.es/)
   - El Mundo (http://www.elmundo.es/)
   - El País (http://elpais.com/)
   - Europapress (http://www.europapress.es/)
   - La Razón (http://www.larazon.es/)
   - La Vanguardia (http://www.lavanguardia.com/)
   - La Voz de Galicia (http://www.lavozdegalicia.es/)
   - RTVE (http://www.rtve.es/)
   - Vozpopuli (http://vozpopuli.com/)

2. Business and economic news sites

   - Economía Digital (http://www.economiadigital.es/)
   - EFE Empresas (http://www.efeempresas.com/)
   - El Economista (http://www.eleconomista.es/)
   - Expansión (http://expansion.com/)
   - Inversión & Finanzas.com (http://www.finanzas.com/)
   - Inbestia (http://inbestia.com/)
   - Libertad Digital (http://www.libertaddigital.com/)
   - Libre Mercado (http://libremercado.com/)
   - Noticias Bancarias (http://noticiasbancarias.com/)

The first (old) iteration of crawling used the following sources. While this list is slightly longer, there are more general news sites, that provide less strictly financial information. The list was revised and a new version was produced, what is called now the "second iteration". We provide the old list for information reason.

1. Newspapers and news services (43 sources in total, 2.793.632 documents):

   - El Pais,
   - El Mundo,
   - ABC,
   - El Confidencial,
   - Yahoo Noticias,
   - Expansion,
   - El Economista,
   - Cinco Dias,
   - Noticias.com,
   - Finanzas.com,
   - Europapress,
   - Economia Digital,
   - EFE Empresas,
   - En Bolsa,
   - Te Interesa,
   - Estrategia de Inversion,
   - La Vanguardia,
   - Bolsamania,

- La Informacion,
- RTVE,
- 20 Minutos,
- Telemadrid Noticias,
- Diario Informacion,
- La Cerca,
- Cadena SER,
- La Voz de Galicia,
- SiliconNews,
- Invertia,
- La Razon,
- Diario Vasco,
- Huffington Post,
- La Opinion Coruna,
- Las Provicias,
- Publico,
- La Rioja,
- FundsPeople,
- El Correo,
- Diario de Leon,
- Dirigentes Digital,
- El Diario Abierto,
- Noticias Bancarias,
- Inbestia,
- El Comercio.

# *B Corpus statistics*

The following pages provide an overview of the corpus statistics. The corpus consiti of news articles, obtained between 1 January 2014 and end of May 2016. It was obrainet through the direct crawling

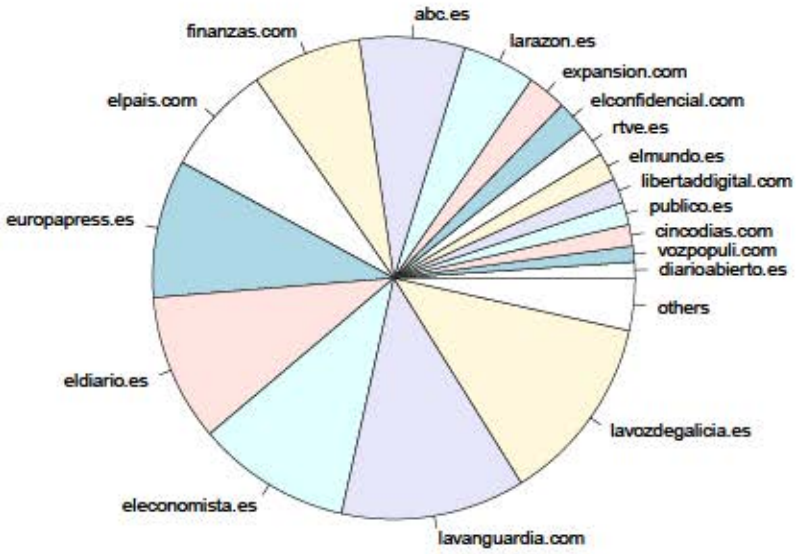



Figure B.1: Volume of acquired news documents per news site as a percentage of total acquired documents. For clarity it shows news sites with more than 20.000 documents

Table B.1: Total number of acquired web documents per news site. The number contains all documents for the given domain and its subdomains.

| | News source | Document count |
|---|---|---|
| 1 | lavozdegalicia.es | 358505 |
| 2 | lavanguardia.com | 349889 |
| 3 | eleconomista.es | 292922 |
| 4 | eldiario.es | 280240 |
| 5 | europapress.es | 262077 |
| 6 | elpais.com | 209686 |
| 7 | finanzas.com | 208268 |
| 8 | abc.es | 200340 |
| 9 | larazon.es | 138429 |
| 10 | expansion.com | 74326 |
| 11 | elconfidencial.com | 59689 |
| 12 | rtve.es | 57691 |
| 13 | elmundo.es | 52781 |
| 14 | libertaddigital.com | 46649 |
| 15 | publico.es | 43141 |
| 16 | cincodias.com | 40393 |
| 17 | vozpopuli.com | 30388 |
| 18 | diarioabierto.es | 30264 |
| 19 | noticiasbancarias.com | 18151 |
| 20 | efeempresas.com | 17389 |
| 21 | economiadigital.es | 16923 |
| 22 | efe.com | 15740 |
| 23 | elcomercio.es | 14401 |
| 24 | libremercado.com | 6028 |
| 25 | inbestia.com | 1267 |

# C Ontology description

This section provides detailed description of the ontologies and taxonomies referred in the Chapter 5.

## C.1 Event class taxonomy

The following table describes classes of the Event taxonomy together with their subclass relations. The common namespace for all concepts is: `http://nadir.uc3m.es/alfredo/events/`.

Table C.1: Event taxonomy class description.

| Name | Definition | Parent class |
|---|---|---|
| Acquisition | Acquisition event describes a situation when a company is participating in an acquisition. The company role is defined through the sub-property of the Role taxonomy. | Transformation |
| AnalystsEarningsEstimates | This event happens when an article is published on future company earnings. There is only one company associated to such event. | AnalystsForecasts |

| Name | Definition | Parent class |
|---|---|---|
| AnalystsForecasts | Super class for all analysts reports concerning the future of a company | FinancialSituationEvent |
| Bankruptcy | This event happens when a bancruptcy of a company is announced. | OtherFinancialSituationEvent |
| BoardCompositionChanges | This event denotes a change of the board of directors of a company. | BoardEvent |
| BoardEvent | This is a superclass containing all events related to the board of a company. | CorporateGovernanceEvent |
| BoardMeetingAnounced | This event takes place when a board meeting is announced. | BoardEvent |
| BuinessExpansion | An event describing a business expansion. | Expansion |
| BusinessScopeChanges | When a company announces an alteration or other modification of its business goals either with a purpose of business expansion, reorganisation or any other reason. | Expansion |
| CEOResignation | An event when a company CEO steps down for whatever reason. | KeyPersonnelEvent |
| ClassAction | A company being a subject of a class action. | LegalProblemEvent |

| Name | Definition | Parent class |
|---|---|---|
| Collaboration | Company announcing a collaboration with other business entity, public body, etc. | StrategicAgreements |
| CompanyDebtManagement | Report is released on company debt management plan. | PeriodicReportEvent |
| CompanyEvent | A super class for all events where a business entity (a company) is a participant. | EventType |
| CompanyRatingChanges | When a new rating for a company is announced. | CompanyRatingEvent |
| CompanyRatingEvent | Super class for company rating events. | FinancialSituationEvent |
| CompanyRestructure | Announcement of company restructure. In principle due to important and transformational reasons, such as reacting to the market needs, changing focus or business priorities, etc. This event is different from mergers and acquisition when restructure also takes place. | Transformation |
| CompanySanctioned | When a company is being sanctioned due to legal problems. | LegalProblemEvent |
| CompanyStatusChange | The change of the status of an business entity. | CompanyStatusEvent |
| CompanyStatusEvent | Super class for any kind of company status change. | CorporateGovernanceEvent |

| Name | Definition | Parent class |
|---|---|---|
| CompanyStocksEvents | Super class for any kind of company stock events | FinancialInstrumentsEvent |
| CompanyUnderInvestigation | Legal investigation against a company is announced. | LegalProblemEvent |
| CompetitionChange | Relevant change in the competition landscape for a business entity. | CompetitionEvent |
| CompetitionEvent | Super class for business competition events. | MarketEnvironmentEvent |
| CorporateGovernanceEvent | A superclass defining all corporate governance events. | CompanyEvent |
| DividendAnnouncement | An announcement of dividend. | CompanyStocksEvents |
| EmplymentEvent | Super class for employment events. | HumanResourcesEvent |
| Expansion | Super class for business expansion events. | StrategicOperationEvent |
| FinancialInstrumentsEvent | Super class for events related to financial instruments emitted by a company (stock). | CompanyEvent |
| FinancialSituationEvent | Super class for all events concerning financial situation of a company. | CompanyEvent |
| FiscalLawIssues | When a fiscal law is being changed or introduced that affects the business or operation of a company. | KeyLawChanges |

| Name | Definition | Parent class |
|---|---|---|
| Fraud | Fraudulent behaviour is revealed. | LegalProblemEvent |
| Hirings | Company announces hirings. | EmplymentEvent |
| HumanResourcesEvent | Super class for all events concerning human resources. Note that this describes only the workforce of a company, not the key personnel. | CompanyEvent |
| IPO | The company IPO is announced. | CompanyStocksEvents |
| JointVenture | A joint-venture kind of partnership is announced. | StrategicAgreements |
| KeyLawChanges | Super class for all events concerning changes in the law that affects company operations. | LegalSituationEvent |
| KeyManagementChange | An event when a change on a key management position is announced. | KeyPersonnelEvent |
| KeyPersonnelEvent | Super class detailing all key personnel changes in a company. | CorporateGovernanceEvent |
| Layoffs | An announcement of layoffs of w company workforce. | EmplymentEvent |
| LegalProblemEvent | Super class for company legal problems. All subclasses of this class describe situations that mostly origin from some sort of misconducts that lead to legal actions. | LegalSituationEvent |

| Name | Definition | Parent class |
|---|---|---|
| LegalSituationEvent | Super class for all events concerning legal issues affecting a company. | CompanyEvent |
| LiquidationAndDissolution | Announcement of termination of a business entity. | Transformation |
| MajorDisaster | A potentially catastrophic event that can lead to major disruptions of a company activity. | OtherExternalEvent |
| MarketEnvironmentEvent | A general class defining all events concerning the market environment of a company. | CompanyEvent |
| Merger | A transaction when a company's ownership changes, typically through a legal consolidation of two or more business entities. | Transformation |
| NewCEOAppointment | An event describing an appointment of a new CEO of a company. | KeyPersonnelEvent |
| NewCompetitor | An event when a new competitor appears on the market. | CompetitionEvent |
| NewProductRelease | A new product is announced to be released on the market. | ProductAndServiceEvent |
| NewStockIssue | A new stock issue is announced (e.g, a secondary offering). | CompanyStocksEvents |
| OtherExternalEvent | A super class for relevant external events not described by other classes of events. | MarketEnvironmentEvent |

| Name | Definition | Parent class |
|---|---|---|
| OtherFinancialSituationEvent | A super class for other important financial situation events not covered by existing classes. | FinancialSituationEvent |
| PatentIssued | When a patent issue is announced. | ProductAndServiceEvent |
| PeriodicReportEvent | Super class for all periodic report announcements. | FinancialSituationEvent |
| ProductAndServiceEvent | Super class for all events related to a company offerings, either products or services. | CompanyEvent |
| ProductIssues | An events when issues with company product are being reported. | ProductAndServiceEvent |
| ProductMarketShareChange | As situation when a flag product's market share changes. | ProductAndServiceEvent |
| QuarterlyReportPublished | A quarterly report is published. | PeriodicReportEvent |
| RawMaterialEvent | Super class for all raw material events. | MarketEnvironmentEvent |
| RawMaterialPriceChange | An event describing a situation when a raw material price changes. | RawMaterialEvent |
| ReverseStockSplit | A reverse stock split is announced. | CompanyStocksEvents |
| SalesOrPerformanceReport | A sales report (or other report concerning company performance) is released. | PeriodicReportEvent |

| Name | Definition | Parent class |
|---|---|---|
| ShareholderRightsChanges | When an important change to the shareholders' rights is announced. | CompanyStatusEvent |
| SpinOff | A company is launching a spin-off entity as a separate business. | Transformation |
| StockBuyback | A buyback is announced. | CompanyStocksEvents |
| StockSplit | A stock split is announced. | CompanyStocksEvents |
| StrategicAgreements | Super class that covers all strategic agreements between business entities. | StrategicOperationEvent |
| StrategicAlliance | An event that describes a situation when a strategic alliance between business entities or organisations is announced. | StrategicAgreements |
| StrategicOperationEvent | Super class for all events concerning strategic decisions regarding either cooperation with other entities or significant change in company operations. | CompanyEvent |
| SubstantialEmployment-Changes | An events describing any other substantial change in the workforce of a company that are not layoffs nor hirings. For instance relocations, subcontracting, etc. | EmplymentEvent |
| TradingSuspension | When company stock trading is being suspended regardless of the reason. | LegalProblemEvent |

| Name | Definition | Parent class |
|---|---|---|
| Transformation | Super class describing all fundamental transformations concerning a business entity. | StrategicOperationEvent |
| YearlyCorporateGovernanceReports | Announcement of periodic corporate governance reports. | CompanyStatusEvent |

## C.2 *Role taxonomy*

This section provide details of the role taxonomy. As pointed out in Section 5.4.1, we are using the EAV-like modelling approach. The following tables describe properties and their attributes. The list has been split into three tables: (i) Table C.2 provide an overview of the top-level relations that are further specialized (ii) Table C.3 describes roles for unary relations (i.e. where only one party participates) and finally (iii) Table C.4 presents roles for binary (or n-ary) relations. Please note that while in the latter case we focus only on binary relations, in some obvious cases we do not restrict the cardinality in order to allow more flexible use in the future. To avoid repetitive information, the cardinality for the domain class (`owl:maxCardinality`) for every property is set to max. 1 unless otherwise stated. The common namespace for all concepts is: `http://nadir.uc3m.es/alfredo/roles/`.

Table C.2: Role taxonomy class description - top-level properties.

| Property Name | Description | Property Axioms |
|---|---|---|
| eventParticipant | This is a top-level property that relates the Event with Entity as a participant in this event. All role properties are sub properties of this one. It is aligned with dbo:participant for possible LOD integration in the future. | **Domain:** Event<br>**Range:** Entity |
| eventSubject | Generic property for describing event subject (i.e. a central entity that governs the event). It is a sub property of eventParticipant. | **Domain:** CompanyEvent<br>**Range:** Company<br>**SubPropertyOf:** eventParticipant |
| eventObject | Generic property for describing event object (i.e. a supplementary entity). It is a sub property of eventParticipant. | **Domain:** CompanyEvent<br>**Range:** AutonomousEntity<br>**SubPropertyOf:** eventParticipant |

Table C.3: Role taxonomy class description - unary relations.

| Property Name | Description | Property Axioms |
|---|---|---|
| analystsEarningsEstimatesForParty | See event class AnalystsEarningsEstimates in Appendix C.1 | **Domain:** AnalystsEarningsEstimates<br>**Range:** Company<br>**SubPropertyOf:** analystsForecastsForParty |

| Property Name | Description | Property Axioms |
|---|---|---|
| analystsForecastForParty | See event class Analysts-Forecasts in Appendix C.1 | **Domain:** AnalystsForecasts<br>**Range:** Company<br>**SubPropertyOf:** financialSituationEventForParty |
| declaredBankruptcy ForParty | See event class Bankruptcy in Appendix C.1 | **Domain:** Bankruptcy<br>**Range:** Company<br>**SubPropertyOf:** otherFinancialSituationEventForParty |
| boardComposition ChangesForParty | See event class Board-CompositionChanges in Appendix C.1 | **Domain:** BoardCompositionChanges<br>**Range:** Company<br>**SubPropertyOf:** boardEventForParty |
| boardEventForParty | See event class BoardEvent in Appendix C.1 | **Domain:** BoardEvent<br>**Range:** Company<br>**SubPropertyOf:** corporateGovernanceEventForParty |
| boardMeeting AnouncedForParty | See event class Board-MeetingAnounced in Appendix C.1 | **Domain:** BoardMeetingAnounced<br>**Range:** Company |

| Property Name | Description | Property Axioms |
|---|---|---|
| | | **SubPropertyOf:** boardEventForParty |
| buinessExpansionForParty | See event class Buiness-Expansion in Appendix C.1 | **Domain:** BuinessExpansion<br>**Range:** Company<br>**SubPropertyOf:** expansionForParty |
| businessScopeChanges ForParty | See event class Busi-nessScopeChanges in Appendix C.1 | **Domain:** BusinessScopeChanges<br>**Range:** Company<br>**SubPropertyOf:** expansionForParty |
| classActionForParty | See event class ClassAc-tion in Appendix C.1 | **Domain:** ClassAction<br>**Range:** Company<br>**SubPropertyOf:** legalProblemEventForParty |
| companyDebtManagement ForParty | See event class Compa-nyDebtManagement in Appendix C.1 | **Domain:** CompanyDebtManagement<br>**Range:** Company<br>**SubPropertyOf:** periodicReportEventForParty |
| companyEventForParty | See event class Com-panyEvent in Appendix C.1 | **Domain:** CompanyEvent<br>**Range:** Company |

| Property Name | Description | Property Axioms |
|---|---|---|
| | | **SubPropertyOf:** eventTypeForParty |
| companyRating ChangesForParty | See event class CompanyRatingChanges in Appendix C.1 | **Domain:** CompanyRatingChanges<br>**Range:** Company<br>**SubPropertyOf:** companyRatingEventForParty |
| companyRating EventForParty | See event class CompanyRatingEvent in Appendix C.1 | **Domain:** CompanyRatingEvent<br>**Range:** Company<br>**SubPropertyOf:** financialSituationEventForParty |
| companyRestructure ForParty | See event class CompanyRestructure in Appendix C.1 | **Domain:** CompanyRestructure<br>**Range:** Company<br>**SubPropertyOf:** transformationForParty |
| companySanctioned ForParty | See event class CompanySanctioned in Appendix C.1 | **Domain:** CompanySanctioned<br>**Range:** Company<br>**SubPropertyOf:** legalProblemEventForParty |
| companyStatusChange ForParty | See event class CompanyStatusChange in Appendix C.1 | **Domain:** CompanyStatusChange |

| Property Name | Description | Property Axioms |
|---|---|---|
| | | **Range:** Company<br>**SubPropertyOf:**<br>companyStatusEventForParty |
| companyStatus EventForParty | See event class CompanyStatusEvent in Appendix C.1 | **Domain:**<br>CompanyStatusEvent<br><br>**Range:** Company<br>**SubPropertyOf:**<br>corporateGovernanceEventForParty |
| companyStocks EventsForParty | See event class CompanyStocksEvents in Appendix C.1 | **Domain:**<br>CompanyStocksEvents<br><br>**Range:** Company<br>**SubPropertyOf:**<br>financialInstrumentsEventForParty |
| companyUnder InvestigationForParty | See event class CompanyUnderInvestigation in Appendix C.1 | **Domain:**<br>CompanyUnderInvestigation<br><br>**Range:** Company<br>**SubPropertyOf:**<br>legalProblemEventForParty |
| competitionChange ForParty | See event class CompetitionChange in Appendix C.1 | **Domain:**<br>CompetitionChange<br><br>**Range:** Company<br>**SubPropertyOf:**<br>competitionEventForParty |

| Property Name | Description | Property Axioms |
|---|---|---|
| competitionEventForParty | See event class CompetitionEvent in Appendix C.1 | **Domain:** CompetitionEvent<br>**Range:** Company<br>**SubPropertyOf:** marketEnvironmentEventForParty |
| corporateGovernanceEventForParty | See event class CorporateGovernanceEvent in Appendix C.1 | **Domain:** CorporateGovernanceEvent<br>**Range:** Company<br>**SubPropertyOf:** companyEventForParty |
| dividendAnnouncementForParty | See event class DividendAnnouncement in Appendix C.1 | **Domain:** DividendAnnouncement<br>**Range:** Company<br>**SubPropertyOf:** companyStocksEventsForParty |
| emplymentEventForParty | See event class EmplymentEvent in Appendix C.1 | **Domain:** EmplymentEvent<br>**Range:** Company<br>**SubPropertyOf:** humanResourcesEventForParty |
| expansionForParty | See event class Expansion in Appendix C.1 | **Domain:** Expansion<br>**Range:** Company |

| Property Name | Description | Property Axioms |
|---|---|---|
| | | **SubPropertyOf:** strategicOperationEventForParty |
| financialInstruments EventForParty | See event class FinancialInstrumentsEvent in Appendix C.1 | **Domain:** FinancialInstrumentsEvent<br>**Range:** Company<br>**SubPropertyOf:** companyEventForParty |
| financialSituation EventForParty | See event class FinancialSituationEvent in Appendix C.1 | **Domain:** FinancialSituationEvent<br>**Range:** Company<br>**SubPropertyOf:** companyEventForParty |
| fiscalLawIssuesForParty | See event class FiscalLawIssues in Appendix C.1 | **Domain:** FiscalLawIssues<br>**Range:** Company<br>**SubPropertyOf:** keyLawChangesForParty |
| fraudForParty | See event class Fraud in Appendix C.1 | **Domain:** Fraud<br>**Range:** Company<br>**SubPropertyOf:** legalProblemEventForParty |
| hiringsForParty | See event class Hirings in Appendix C.1 | **Domain:** Hirings<br>**Range:** Company |

| Property Name | Description | Property Axioms |
|---|---|---|
| | | **SubPropertyOf:** emplymentEventForParty |
| humanResources EventForParty | See event class HumanResourcesEvent in Appendix C.1 | **Domain:** HumanResourcesEvent<br>**Range:** Company<br>**SubPropertyOf:** companyEventForParty |
| ipoForParty | See event class Ipo in Appendix C.1 | **Domain:** Ipo<br>**Range:** Company<br>**SubPropertyOf:** companyStocksEventsForParty |
| keyLawChangesForParty | See event class KeyLawChanges in Appendix C.1 | **Domain:** KeyLawChanges<br>**Range:** Company<br>**SubPropertyOf:** legalSituationEventForParty |
| keyManagement ChangeForParty | See event class KeyManagementChange in Appendix C.1 | **Domain:** KeyManagementChange<br>**Range:** Company<br>**SubPropertyOf:** keyPersonnelEventForParty |
| keyPersonnelEvent ForParty | See event class KeyPersonnelEvent in Appendix C.1 | **Domain:** KeyPersonnelEvent<br>**Range:** Company |

| Property Name | Description | Property Axioms |
|---|---|---|
| | | **SubPropertyOf:** corporateGovernanceEventForParty |
| layoffsForParty | See event class Layoffs in Appendix C.1 | **Domain:** Layoffs **Range:** Company **SubPropertyOf:** emplymentEventForParty |
| legalProblemEvent ForParty | See event class Legal-ProblemEvent in Appendix C.1 | **Domain:** LegalProblemEvent **Range:** Company **SubPropertyOf:** legalSituationEventForParty |
| legalSituationEvent ForParty | See event class Legal-SituationEvent in Appendix C.1 | **Domain:** LegalSituationEvent **Range:** Company **SubPropertyOf:** companyEventForParty |
| liquidationAnd DissolutionForParty | See event class Liquida-tionAndDissolution in Appendix C.1 | **Domain:** LiquidationAndDissolution **Range:** Company **SubPropertyOf:** transformationForParty |
| majorDisasterForParty | See event class MajorDis-aster in Appendix C.1 | **Domain:** MajorDisaster **Range:** Company |

| Property Name | Description | Property Axioms |
|---|---|---|
| | | **SubPropertyOf:** otherExternalEventForParty |
| marketEnvironment EventForParty | See event class MarketEnvironmentEvent in Appendix C.1 | **Domain:** MarketEnvironmentEvent<br>**Range:** Company<br>**SubPropertyOf:** companyEventForParty |
| newMarketExpansion ForParty | See event class NewMarketExpansion in Appendix C.1 | **Domain:** NewMarketExpansion<br>**Range:** Company<br>**SubPropertyOf:** expansionForParty |
| newStockIssueForParty | See event class NewStockIssue in Appendix C.1 | **Domain:** NewStockIssue<br>**Range:** Company<br>**SubPropertyOf:** companyStocksEventsForParty |
| otherExternalEvent ForParty | See event class OtherExternalEvent in Appendix C.1 | **Domain:** OtherExternalEvent<br>**Range:** Company<br>**SubPropertyOf:** marketEnvironmentEventForParty |
| otherFinancialSituation EventForParty | See event class OtherFinancialSituationEvent in Appendix C.1 | **Domain:** OtherFinancialSituationEvent |

| Property Name | Description | Property Axioms |
|---|---|---|
| | | **Range:** Company<br>**SubPropertyOf:**<br>financialSituationEventForParty |
| patentIssuedForParty | See event class PatentIssued in Appendix C.1 | **Domain:**<br>PatentIssued<br>**Range:** Company<br>**SubPropertyOf:**<br>productAndServiceEventForParty |
| periodicReport EventForParty | See event class PeriodicReportEvent in Appendix C.1 | **Domain:**<br>PeriodicReportEvent<br>**Range:** Company<br>**SubPropertyOf:**<br>financialSituationEventForParty |
| productMarketShare ChangesForParty | See event class ProductMarketShareChanges in Appendix C.1 | **Domain:**<br>ProductMarketShareChanges<br>**Range:** Company<br>**SubPropertyOf:**<br>productAndServiceEventForParty |
| quarterlyReport PublishedForParty | See event class QuarterlyReportPublished in Appendix C.1 | **Domain:**<br>QuarterlyReportPublished<br>**Range:** Company<br>**SubPropertyOf:**<br>periodicReportEventForParty |
| reverseStockSplitForParty | See event class ReverseStockSplit in Appendix C.1 | **Domain:**<br>ReverseStockSplit |

| Property Name | Description | Property Axioms |
|---|---|---|
| | | **Range:** Company<br>**SubPropertyOf:**<br>companyStocksEventsForParty |
| salesOrPerformance ReportForParty | See event class SalesOr-PerformanceReport in Appendix C.1 | **Domain:**<br>SalesOrPerformanceReport<br><br>**Range:** Company<br>**SubPropertyOf:**<br>periodicReportEventForParty |
| sectorEventForParty | See event class Sec-torEvent in Appendix C.1 | **Domain:**<br>SectorEvent<br><br>**Range:** Company<br>**SubPropertyOf:**<br>marketEnvironmentEventForParty |
| shareholderRights ChangesForParty | See event class Share-holderRightsChanges in Appendix C.1 | **Domain:**<br>ShareholderRightsChanges<br><br>**Range:** Company<br>**SubPropertyOf:**<br>companyStatusEventForParty |
| spinOffForParty | See event class SpinOff in Appendix C.1 | **Domain:**<br>SpinOff<br>**Range:** Company<br>**SubPropertyOf:**<br>transformationForParty |
| stockBuybackForParty | See event class Stock-Buyback in Appendix C.1 | **Domain:**<br>StockBuyback |

| Property Name | Description | Property Axioms |
|---|---|---|
| | | **Range:** Company<br>**SubPropertyOf:**<br>companyStocksEventsForParty |
| stockSplitForParty | See event class StockSplit in Appendix C.1 | **Domain:**<br>StockSplit<br>**Range:** Company<br>**SubPropertyOf:**<br>companyStocksEventsForParty |
| strategicAgreements ForParty | See event class StrategicAgreements in Appendix C.1 | **Domain:**<br>StrategicAgreements<br>**Range:** Company<br>**SubPropertyOf:**<br>strategicOperationEventForParty |
| strategicAllianceForParty | See event class StrategicAlliance in Appendix C.1 | **Domain:**<br>StrategicAlliance<br>**Range:** Company<br>**SubPropertyOf:**<br>strategicAgreementsForParty |
| strategicOperation EventForParty | See event class StrategicOperationEvent in Appendix C.1 | **Domain:**<br>StrategicOperationEvent<br>**Range:** Company<br>**SubPropertyOf:**<br>companyEventForParty |

| Property Name | Description | Property Axioms |
|---|---|---|
| substantialEmploymentChangesForParty | See event class SubstantialEmploymentChanges in Appendix C.1 | **Domain:** SubstantialEmploymentChanges<br>**Range:** Company<br>**SubPropertyOf:** emplymentEventForParty |
| tradingSuspensionForParty | See event class TradingSuspension in Appendix C.1 | **Domain:** TradingSuspension<br>**Range:** Company<br>**SubPropertyOf:** legalProblemEventForParty |
| transformationForParty | See event class Transformation in Appendix C.1 | **Domain:** Transformation<br>**Range:** Company<br>**SubPropertyOf:** strategicOperationEventForParty |
| yearlyCorporateGovernanceReportsForParty | See event class YearlyCorporateGovernanceReports in Appendix C.1 | **Domain:** YearlyCorporateGovernanceReports<br>**Range:** Company<br>**SubPropertyOf:** companyStatusEventForParty |

Table C.4: Role taxonomy class description - binary relations.

| Property Name | Description | Property Axioms |
|---|---|---|
| ceoAppointingCompany | This property describes a Company role in a NewCEOAppointment event. That is: a company appointing its new CEO. | **Domain:** NewCEOAppointed<br>**Range:** Company<br>**SubPropertyOf:** eventSubject |
| appointedCeo | This property describes a Person role in a NewCEOAppointment event, i.e.: a Person that becomes a new CEO. | **Domain:** NewCEOAppointed<br>**Range:** Person<br>**SubPropertyOf:** eventObject |
| ceoResignsFromCompany | This property describes a Company role in a CEOResignation event. That is: a company where the CEO resigned. | **Domain:** CEOResignation<br>**Range:** Company<br>**SubPropertyOf:** eventSubject |
| resigningCeo | This property describes a Person role in a CEOResignation event, i.e.: a Person that resignes (or is forced to do so) from the CEO role. | **Domain:** CEOResignation<br>**Range:** Person<br>**SubPropertyOf:** eventObject |
| spinoffParentCompany | This property describes a parent company launching a new Spin-Off | **Domain:** SpinOff<br>**Range:** Company<br>**SubPropertyOf:** eventSubject |

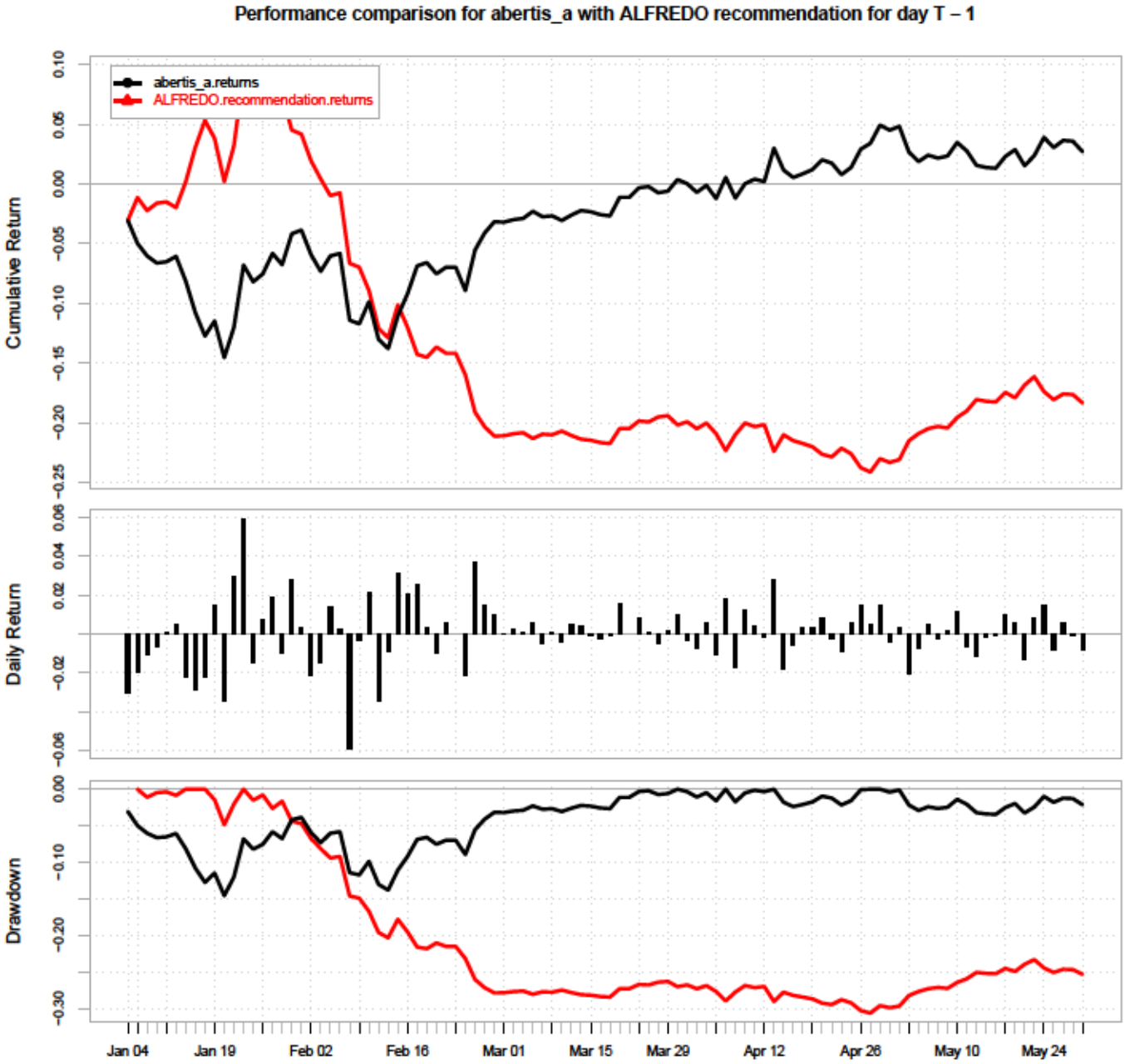
| Property Name | Description | Property Axioms |
|---|---|---|
| spinoffNewCompany | This property describes a new company that is the actual Spin-Off. | **Domain:** SpinOff<br>**Range:** Company<br>**SubPropertyOf:**<br>eventObject |
| productReleasing Company | This property describes a Company role in a generic event of class ProductAndServiceEvent. | **Domain:** ProductAndServiceEvent<br>**Range:** Company<br>**SubPropertyOf:**<br>eventSubject |
| productName | This property describes a ProductOrService role in a generic event of class ProductAndServiceEvent. Note that this property need to inherit from the top-level property, due to the fact that products are not subclass of AutonomousEntity. | **Domain:** ProductAndServiceEvent<br>**Range:** ProductOrService<br>**SubPropertyOf:**<br>eventParticipant |
| newlyReleased ProductCompany | This property describes a Company role in a event NewProductRelease, i.e. It points to a company entity that launches a new product. | **Domain:** NewProductRelease<br>**Range:** Company<br>**SubPropertyOf:**<br>eventSubject |
| newlyReleased ProductName | This property describes a ProductOrService role in a event NewProductRelease, i.e. It points to a new product (or service) that is launched by a company. | **Domain:** NewProductRelease<br>**Range:** ProductOrService<br>**SubPropertyOf:**<br>eventParticipant |

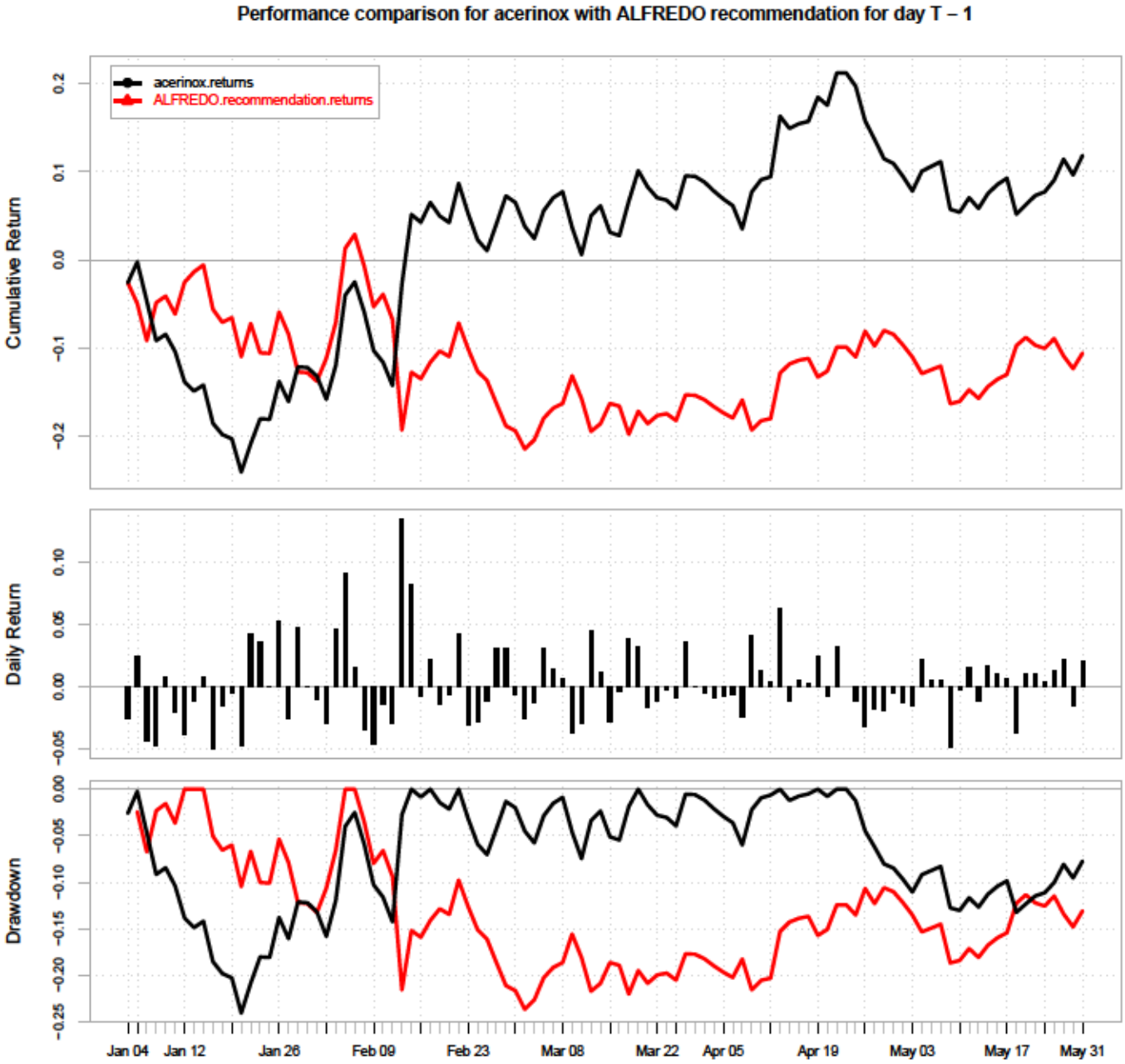
| Property Name | Description | Property Axioms |
|---|---|---|
| productWithIssues Company | This property describes a Company role in a event ProductIssues, i.e. It points to a company entity that has issues with its product. | **Domain:** ProductIssues **Range:** Company **SubPropertyOf:** eventSubject |
| productWithIssues Name | This property describes a ProductOrService role in a event ProductIssues, i.e. It points to a new product (or service) that is a subject of issues. | **Domain:** ProductIssues **Range:** ProductOrService **SubPropertyOf:** eventParticipant |
| establishedCompany | This property describes a Company role in a event NewCompetitor, i.e. It points to an established company that faces a new competitor. | **Domain:** NewCompetitor **Range:** Company **SubPropertyOf:** eventSubject |
| newCompetitorCompany | This property describes a Company role in a event NewCompetitor, i.e. It points to a company entity that is the new competition. | **Domain:** NewCompetitor **Range:** Company **SubPropertyOf:** eventObject |
| mergingCompany | This property describes a Company role in a event Merger. Note that we do not distinguish between merging entities, thus the same property is used for each merger participant. Although we focus on binary events, we do not limit the cardinality of this relation. | **Domain:** Merger **Range:** Company **SubPropertyOf:** eventSubject |

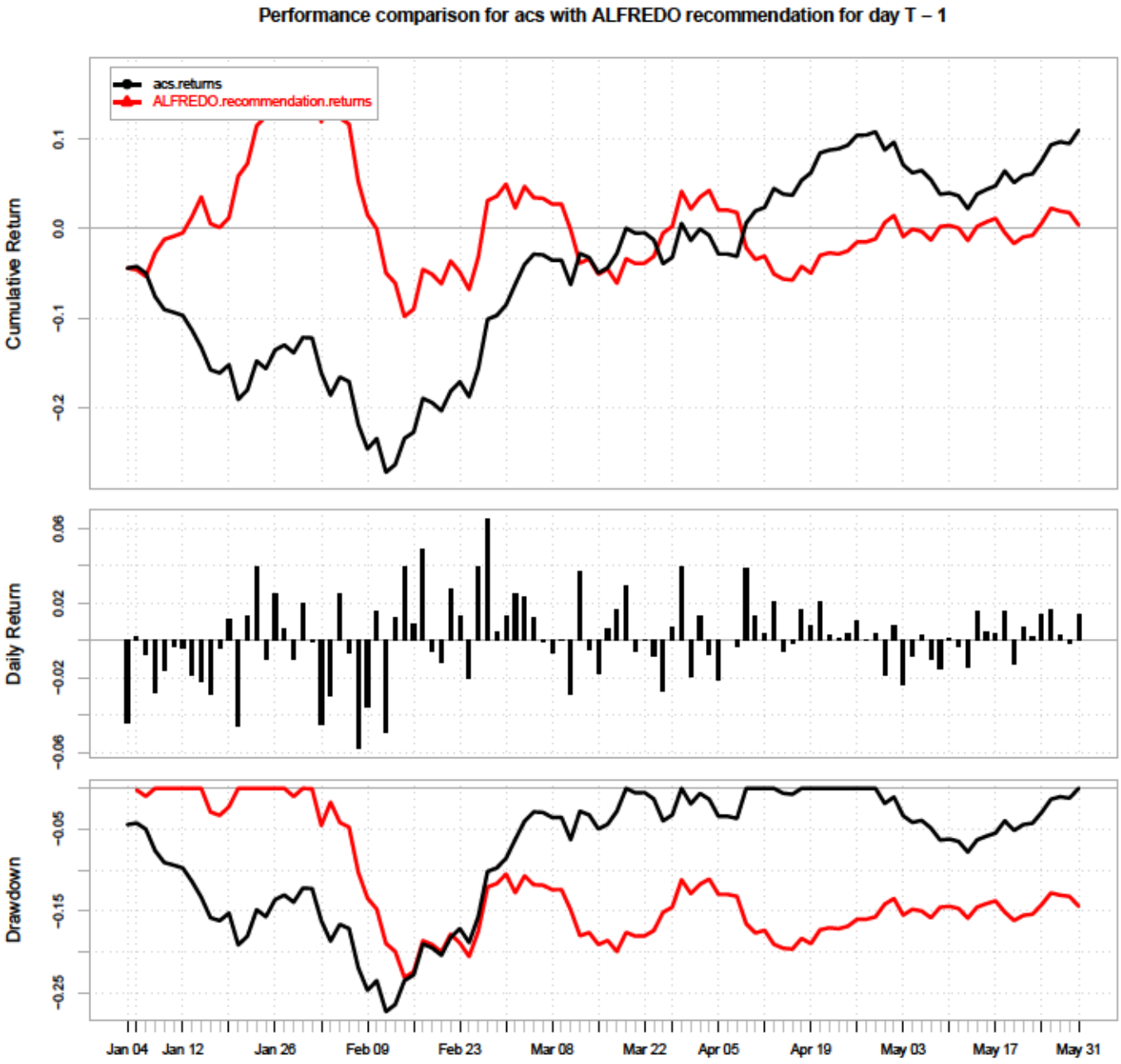
| Property Name | Description | Property Axioms |
|---|---|---|
| jointVenturePartner Company | This property describes a Company role in a event JointVenture. Note that we do not distinguish between joint-venture entities, thus the same property is used for each participant. Although we focus on binary events, we do not limit the cardinality of this relation. | **Domain:** JointVenture **Range:** Company **SubPropertyOf:** eventSubject |
| collaborationPartner Company | This property describes a Company role in a event Collaboration. Note that we do not distinguish between collaborating entities, thus the same property is used for each participant. Although we focus on binary events, we do not limit the cardinality of this relation. | **Domain:** Collaboration **Range:** Company **SubPropertyOf:** eventSubject |

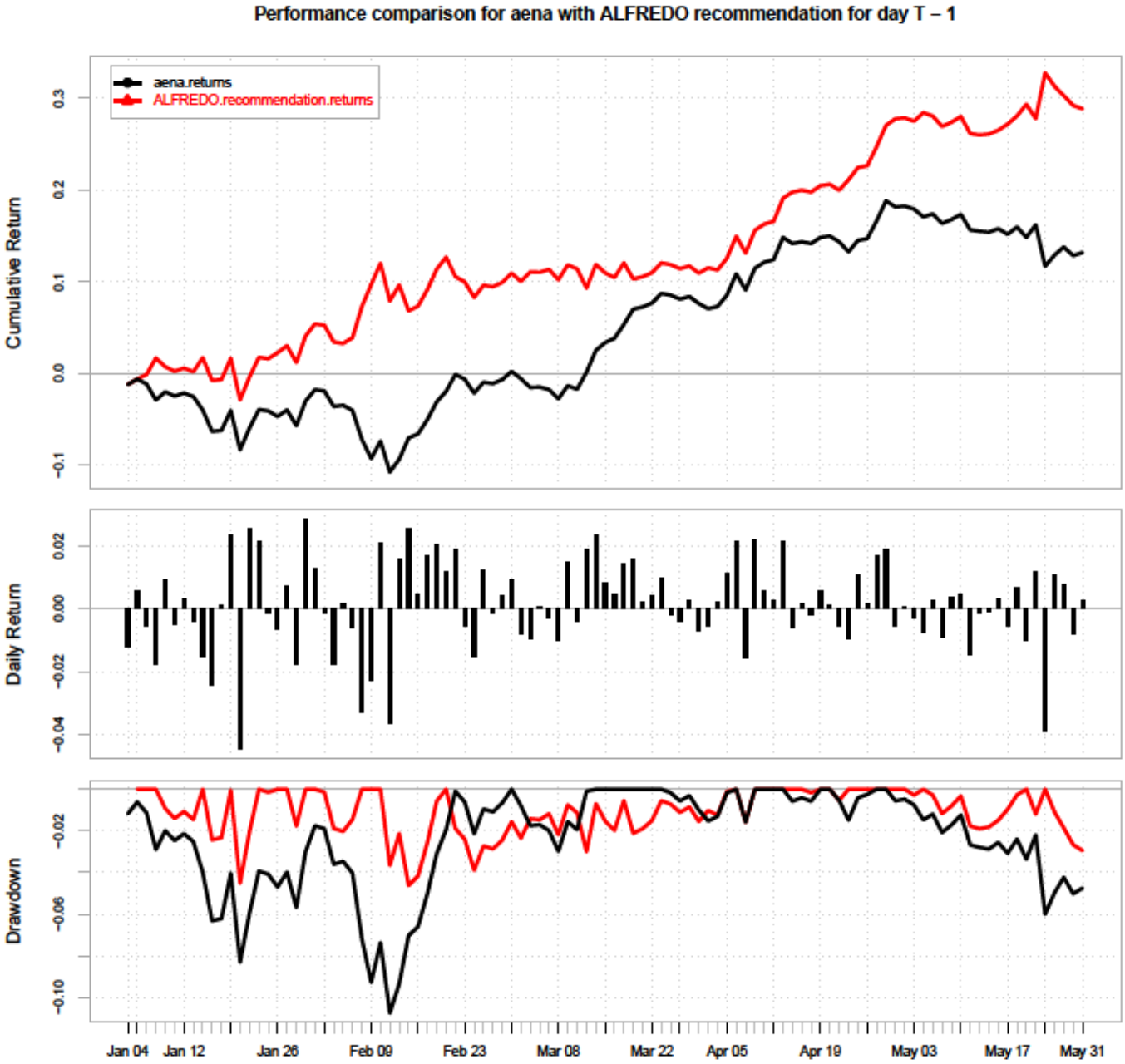
# *D Detailed evaluation results*

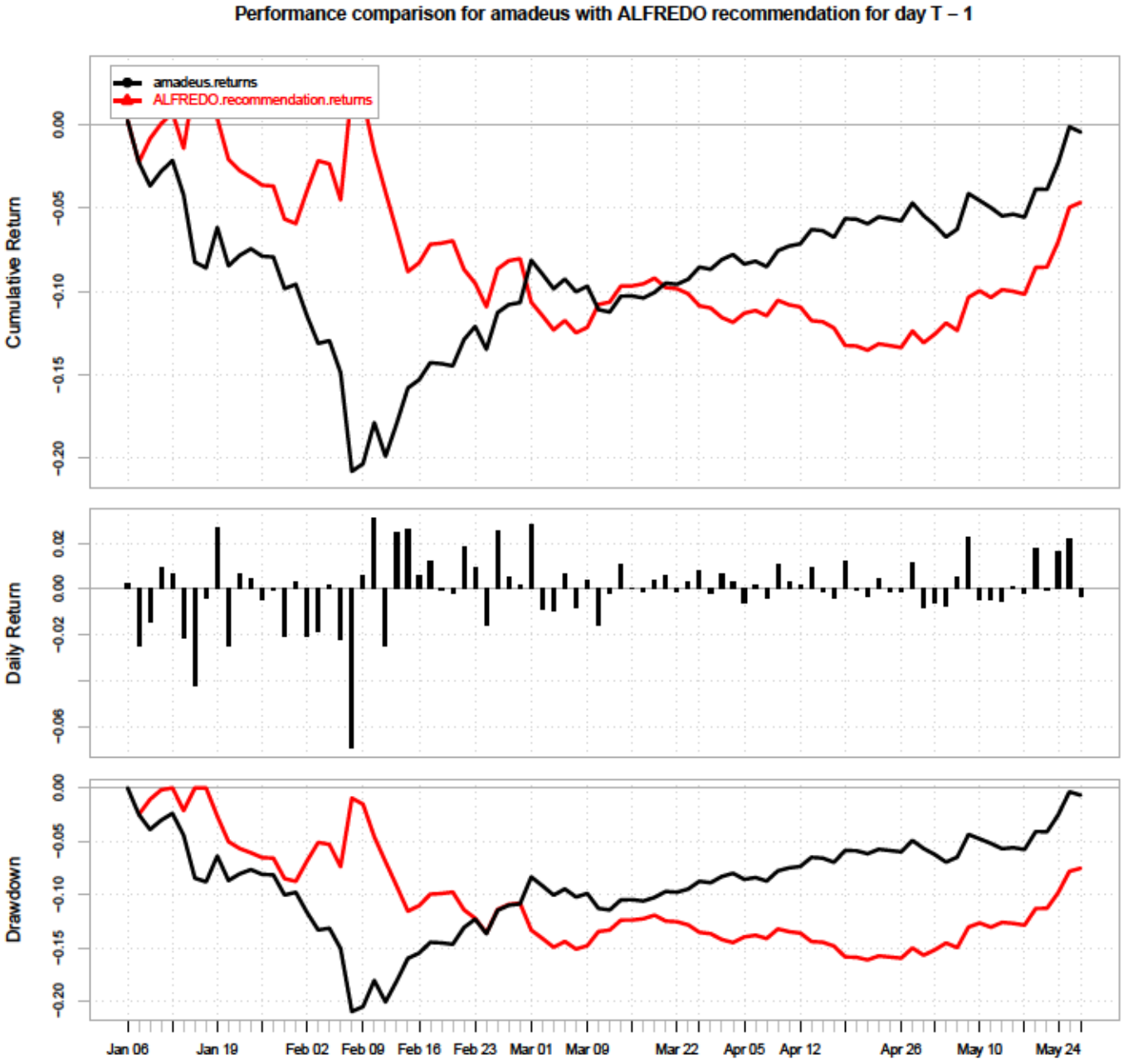
On the following pages you will find cumulative return charts as a result of ALFREDO backtesting in the $T$ 1 scenario against the IBEX35 companies. Each chart contains 3 parts: the upper chart shows cumulative returns, the middle bar chart shows daily returns and the bottom chart shows *drawdowns*. The drawdowns is the difference between latest peak value and the subsequent trough. High drawdowns are typically associated with high risks.

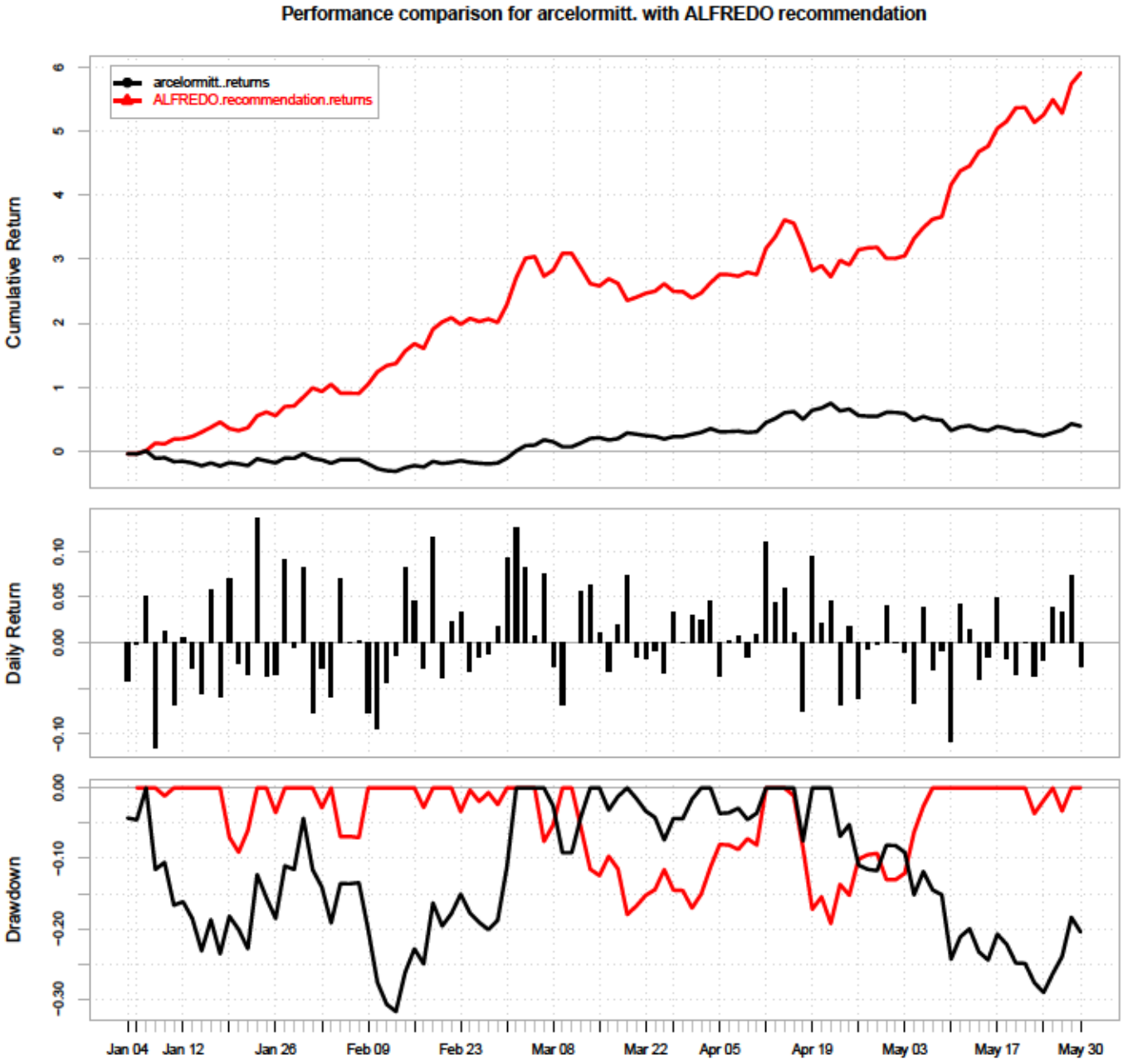

Performance comparison for abertis_a with ALFREDO recommendation for day T – 1
abertis_a.returns
ALFREDO.recommendation.returns
Cumulative Return
Daily Return
Drawdown
Jan 04
Jan 19
Feb 02
Feb 16
Mar 01
Mar 15
Mar 29
Apr 12
Apr 26
May 10
May 24

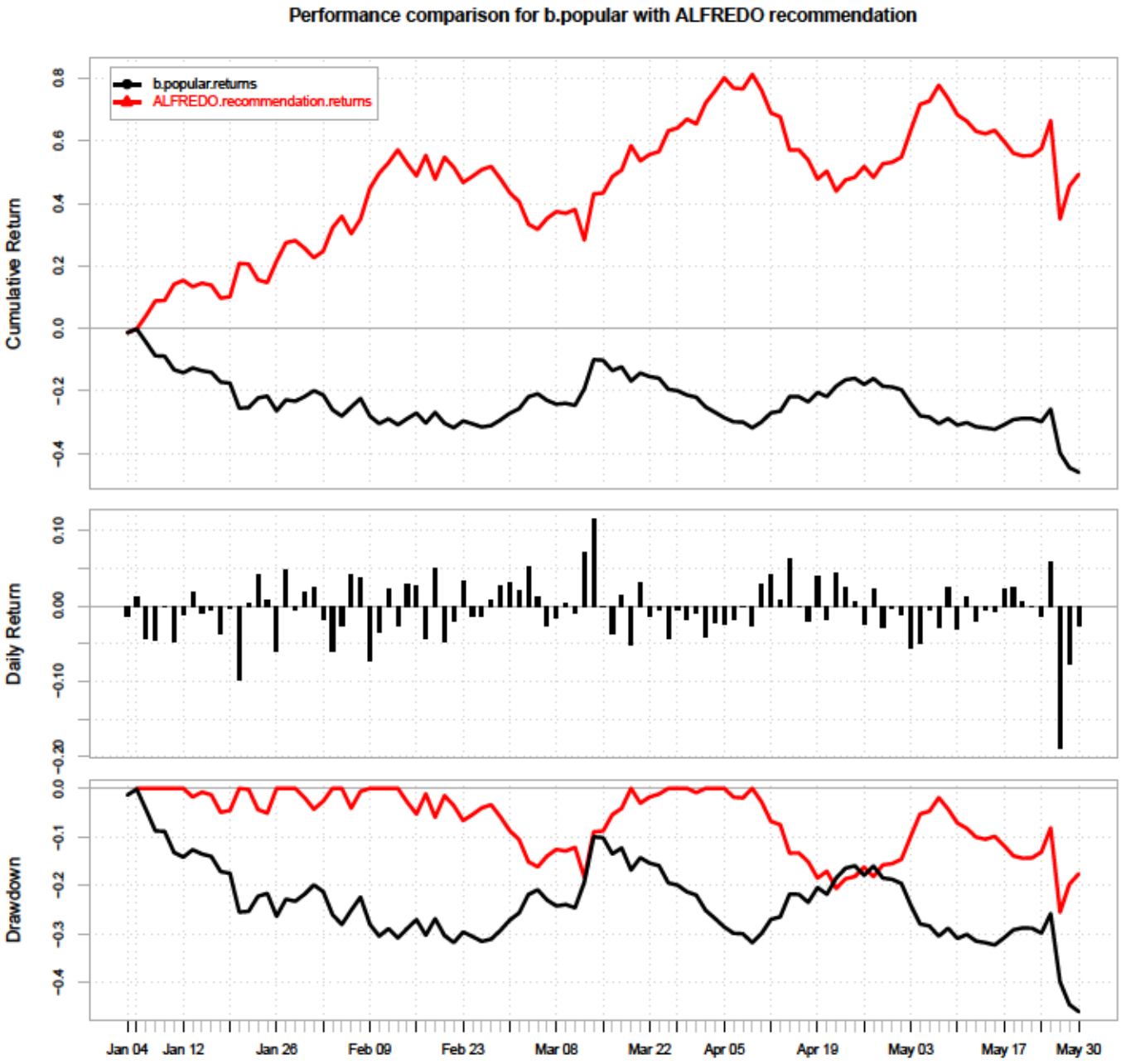

Performance comparison for acerinox with ALFREDO recommendation for day T – 1
acerinox.returns
ALFREDO.recommendation.returns
Cumulative Return
Daily Return
Drawdown
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31

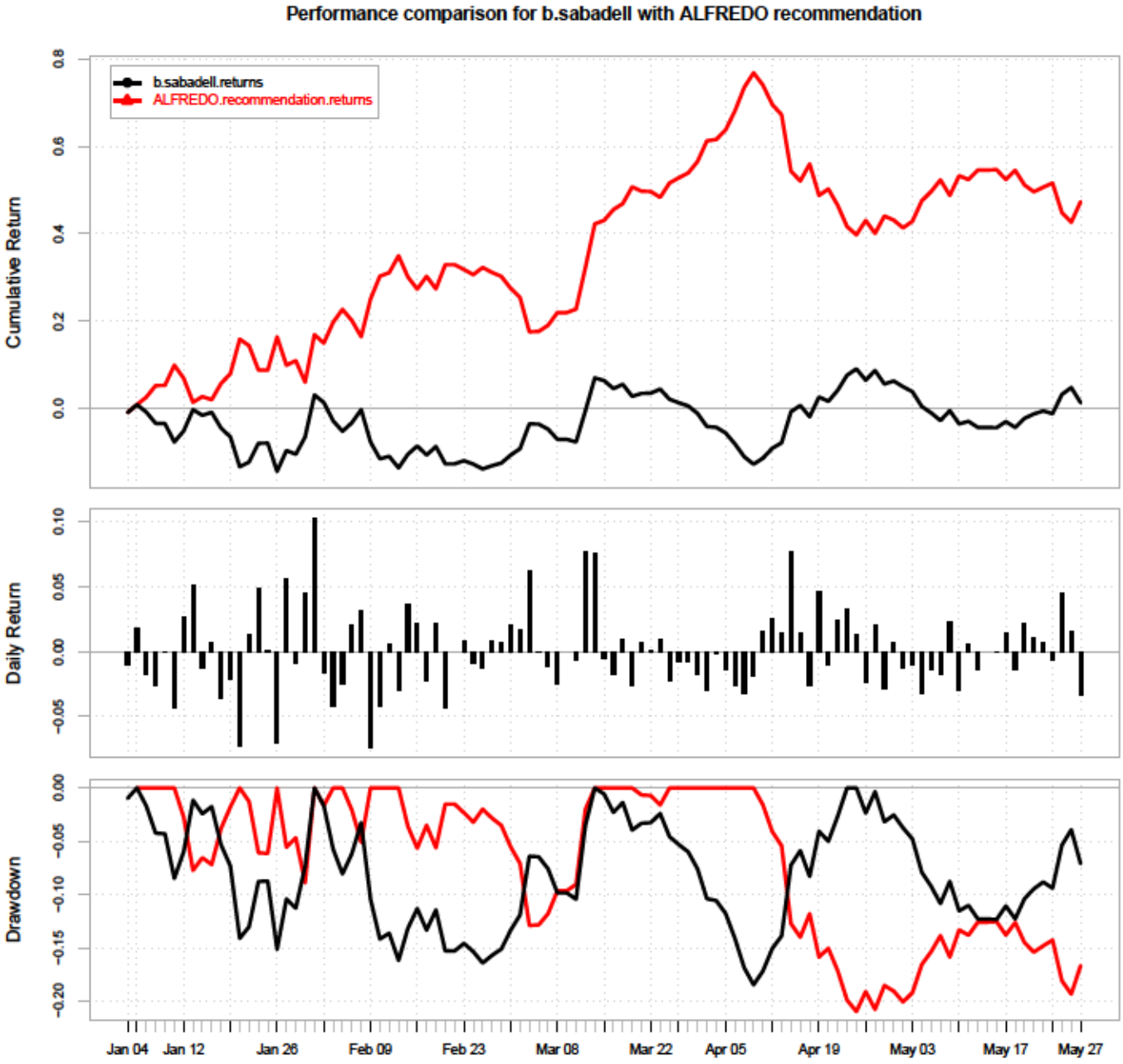

Performance comparison for acs with ALFREDO recommendation for day T – 1
acs.returns
ALFREDO.recommendation.returns
Cumulative Return
0.1
0.0
−0.1
−0.2
Daily Return
0.06
0.02
−0.02
−0.06
Drawdown
−0.05
−0.15
−0.25
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31

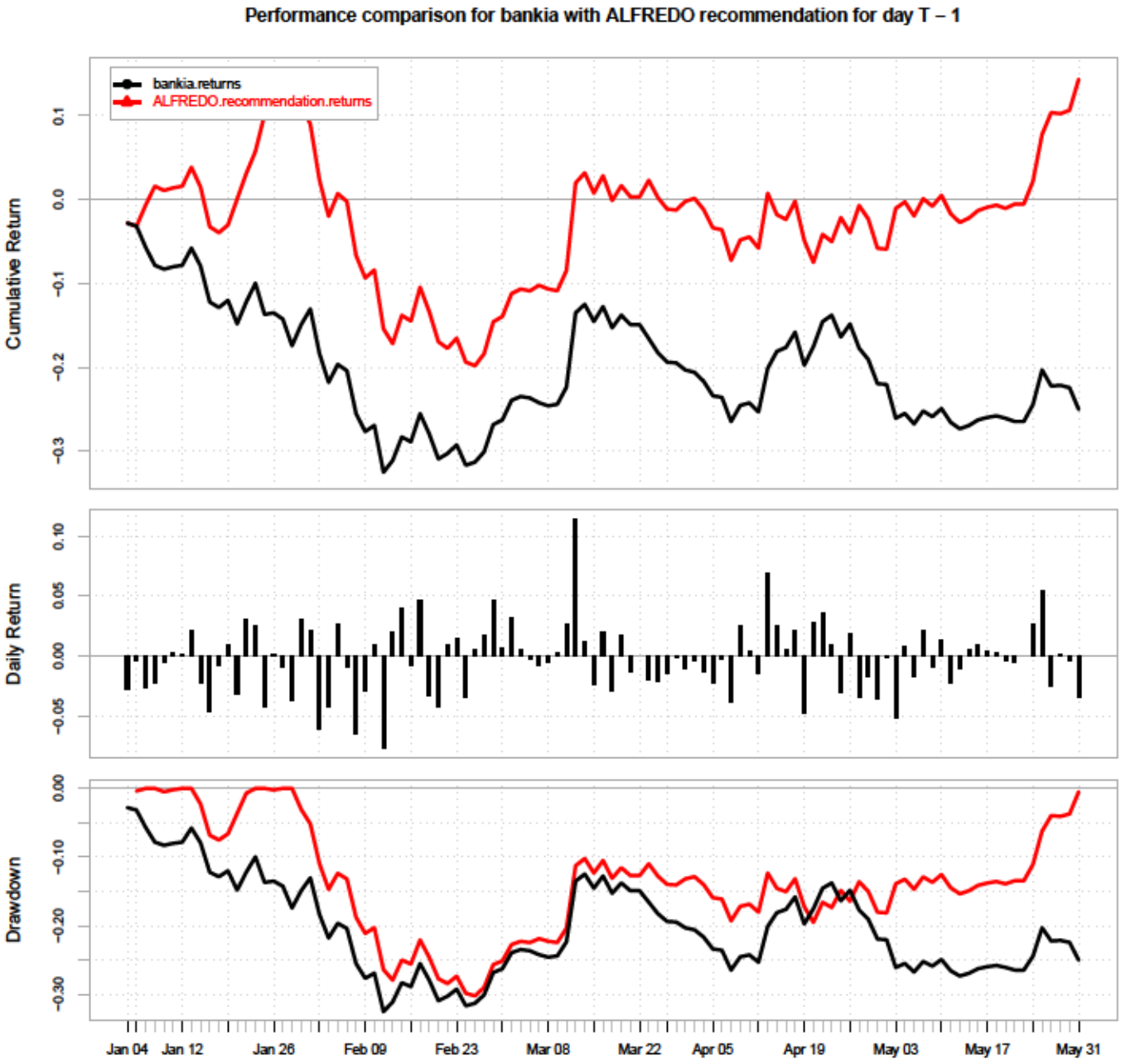

Performance comparison for aena with ALFREDO recommendation for day T – 1
aena.returns
ALFREDO.recommendation.returns
Cumulative Return
0.3
0.2
0.1
0.0
-0.1
Daily Return
0.02
0.00
-0.02
-0.04
Drawdown
-0.02
-0.06
-0.10
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31

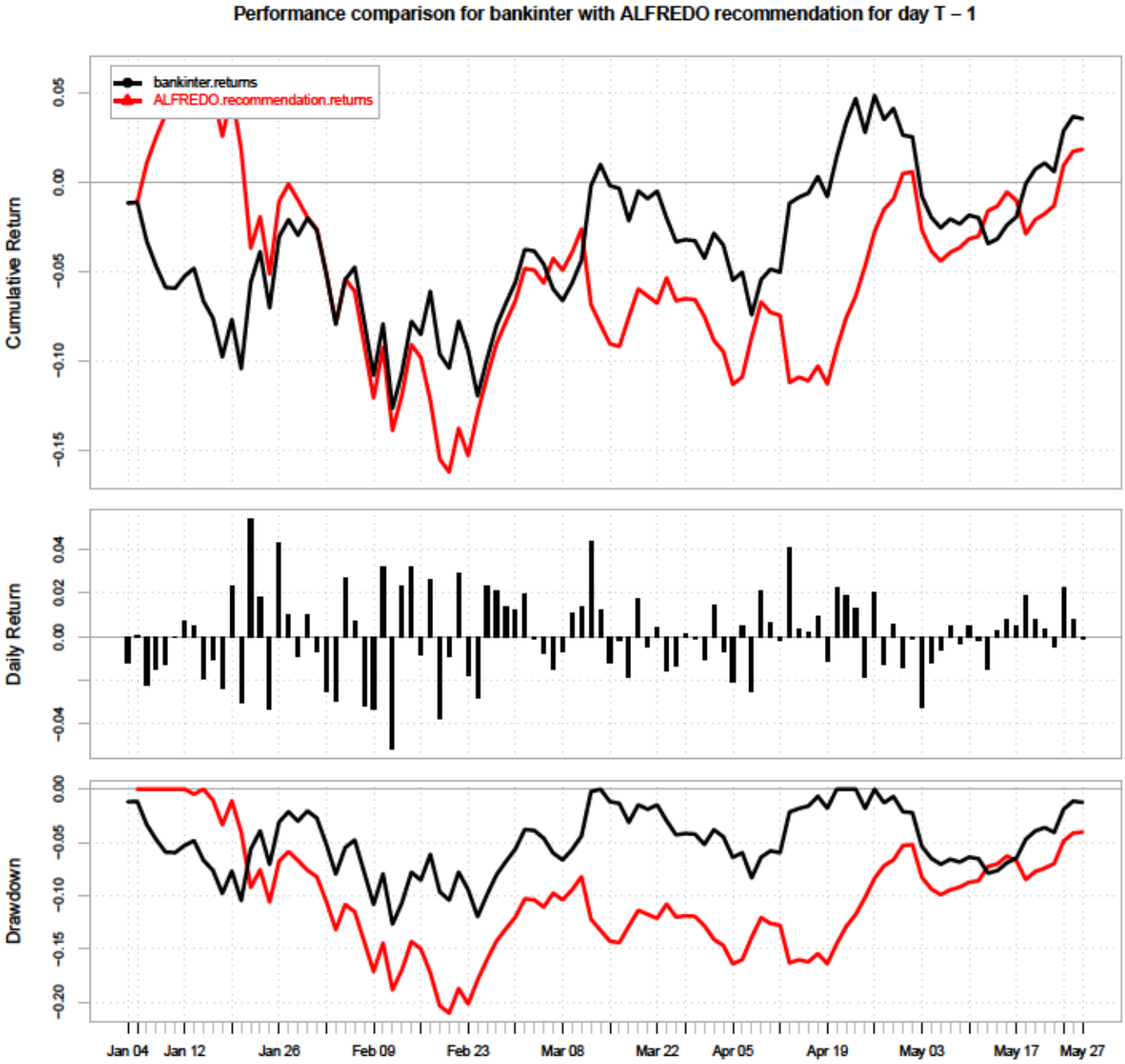

Performance comparison for amadeus with ALFREDO recommendation for day T – 1
amadeus.returns
ALFREDO.recommendation.returns
Cumulative Return
0.00
-0.05
-0.10
-0.15
-0.20
Daily Return
0.02
0.00
-0.02
-0.06
Drawdown
0.00
-0.05
-0.10
-0.15
-0.20
Jan 06
Jan 19
Feb 02
Feb 09
Feb 16
Feb 23
Mar 01
Mar 09
Mar 22
Apr 05
Apr 12
Apr 26
May 10
May 24

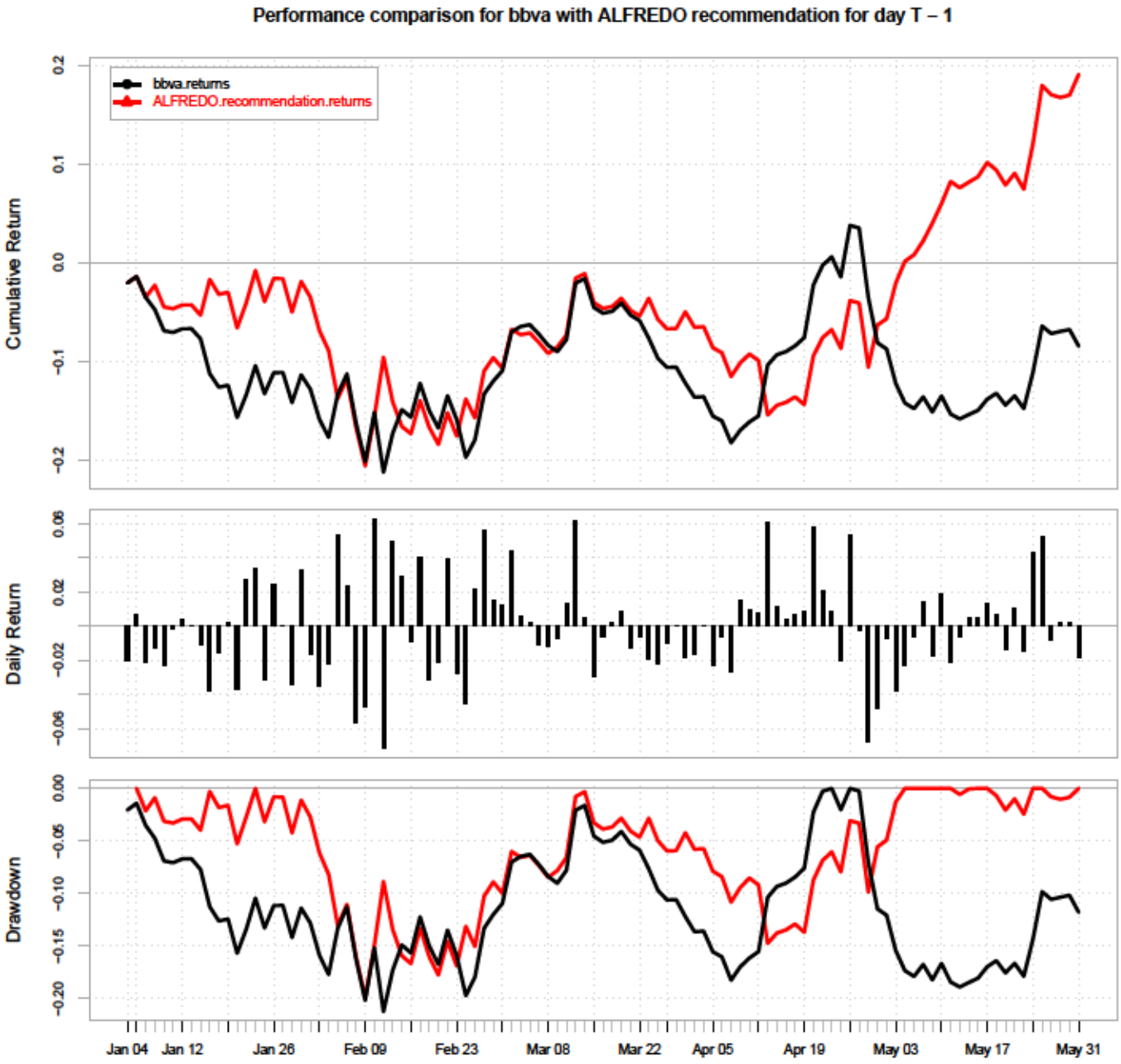

Performance comparison for arcelormitt. with ALFREDO recommendation
arcelormitt..returns
ALFREDO.recommendation.returns
Cumulative Return
Daily Return
Drawdown
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 30

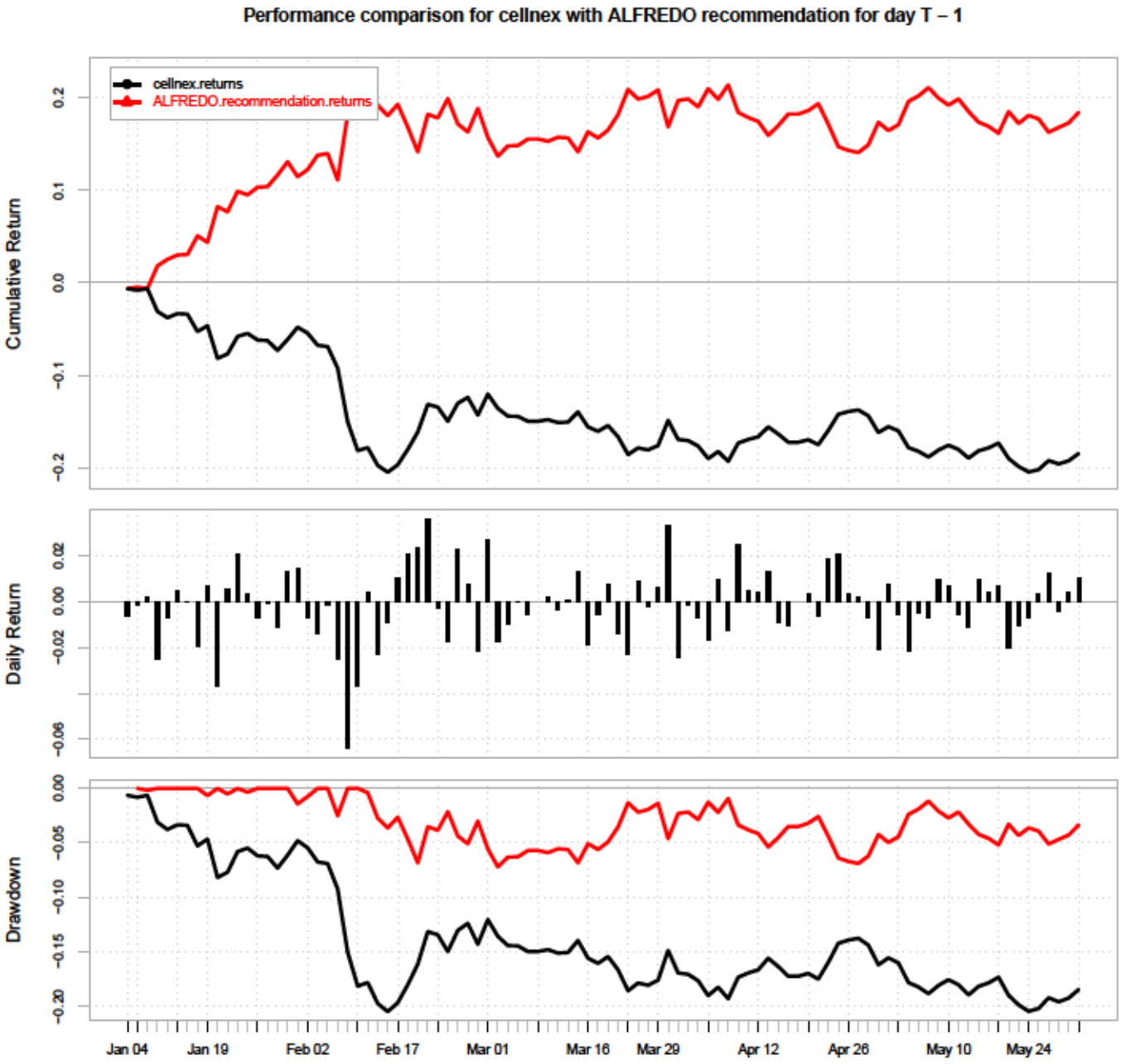

Performance comparison for b.popular with ALFREDO recommendation
b.popular.returns
ALFREDO.recommendation.returns
Cumulative Return
Daily Return
Drawdown
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 30

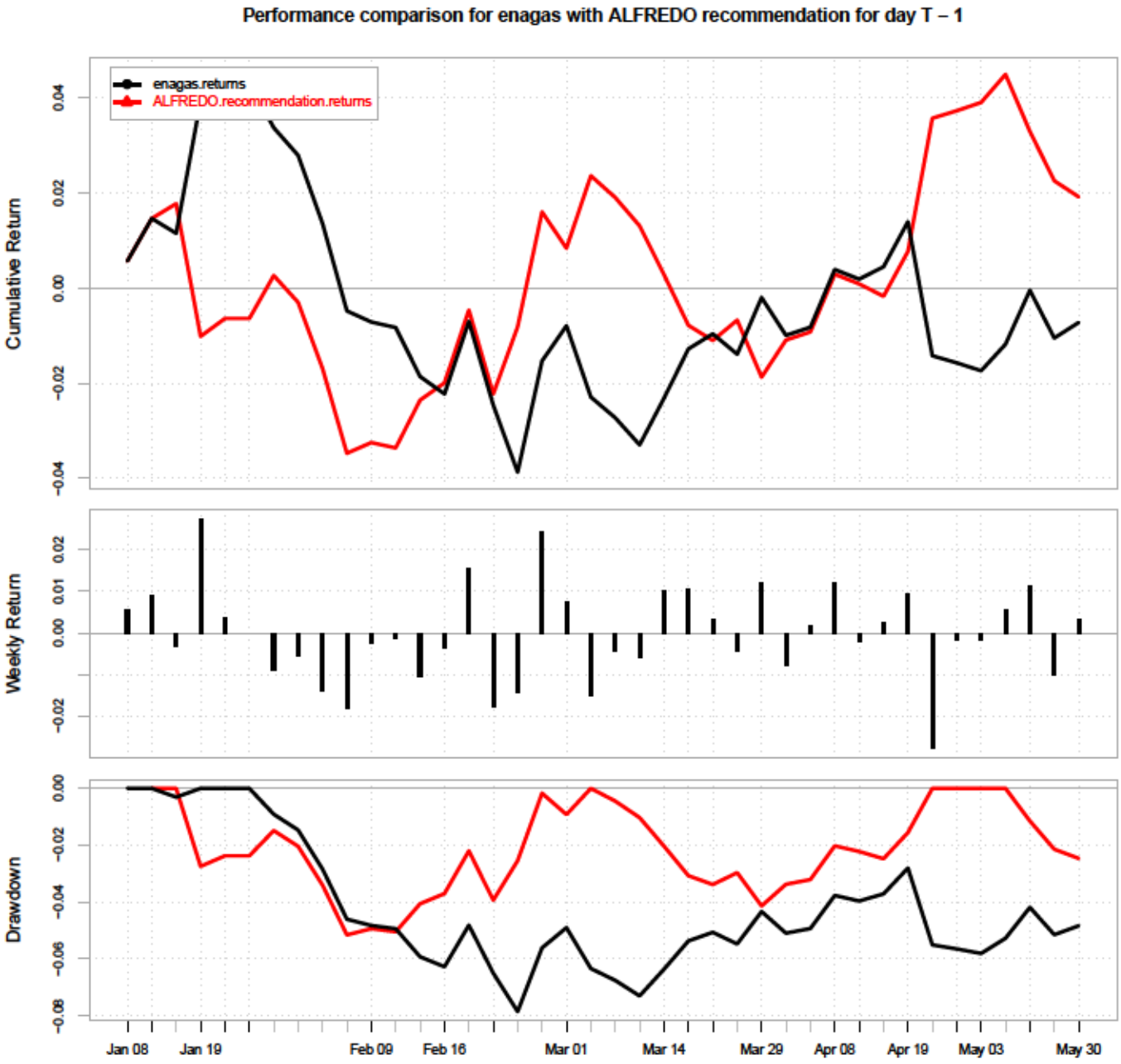

Performance comparison for b.sabadell with ALFREDO recommendation
b.sabadell.returns
ALFREDO.recommendation.returns
Cumulative Return
Daily Return
Drawdown
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 27

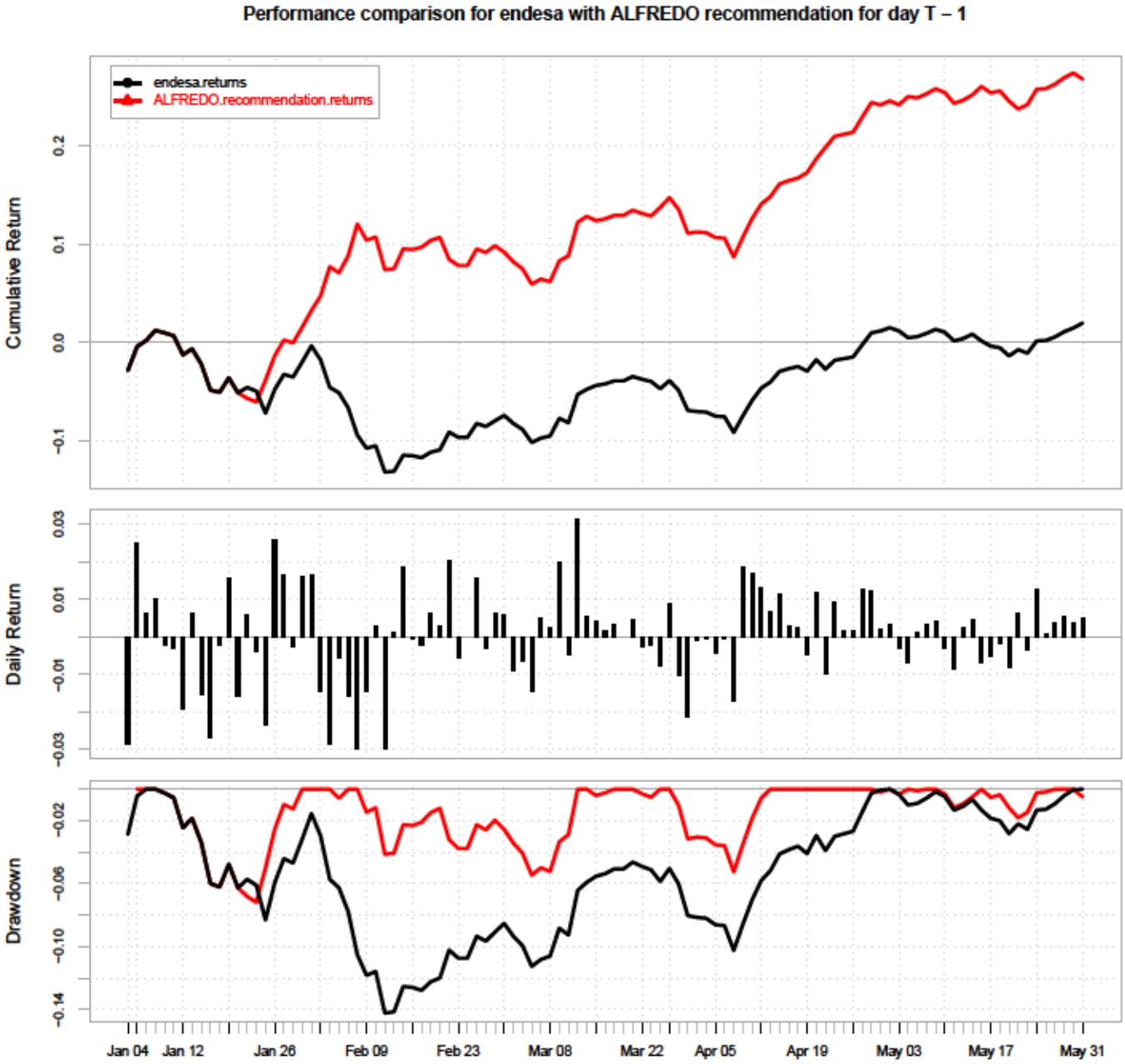

Performance comparison for bankia with ALFREDO recommendation for day T – 1
bankia.returns
ALFREDO.recommendation.returns
Cumulative Return
Daily Return
Drawdown
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31

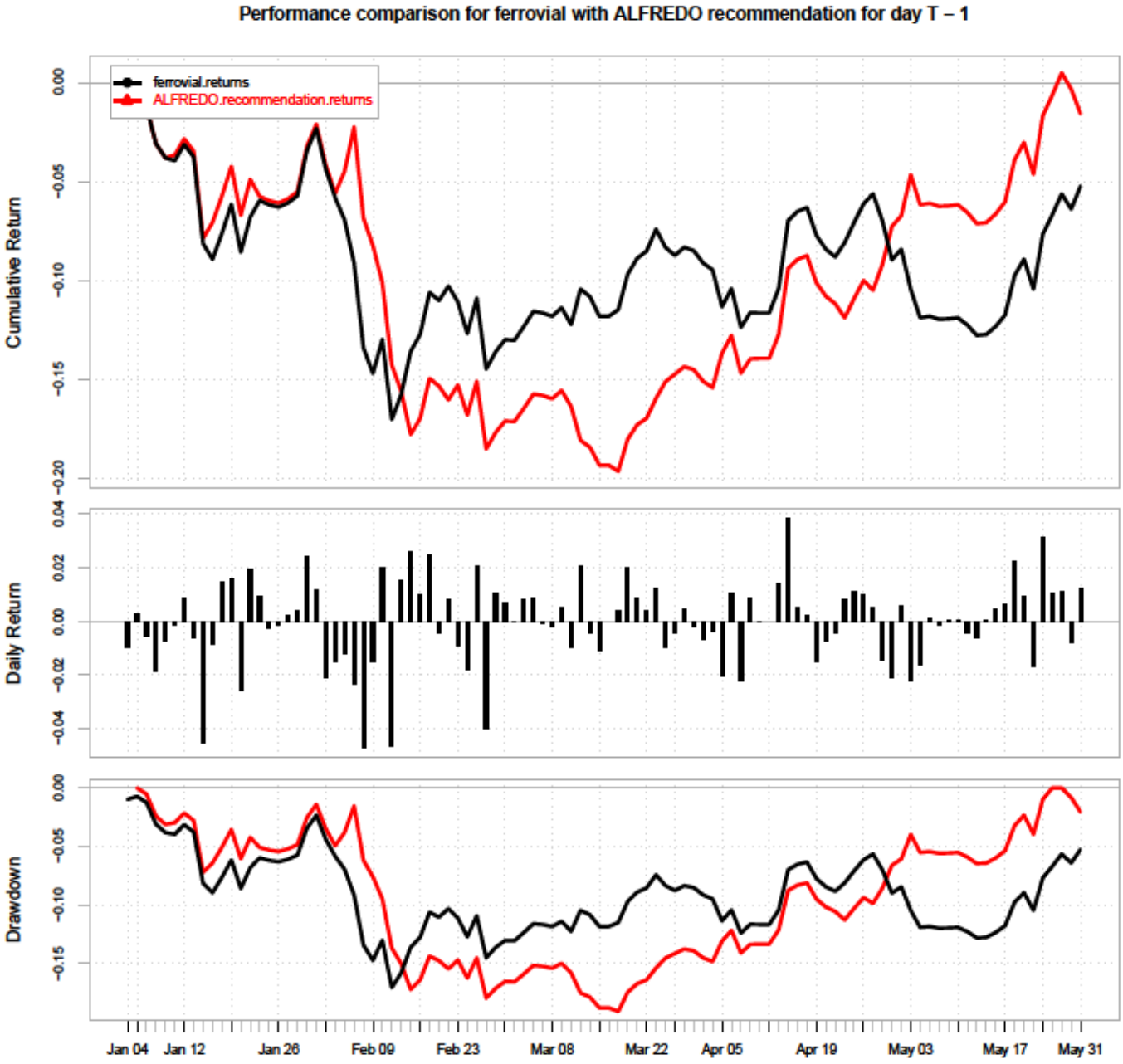

Performance comparison for bankinter with ALFREDO recommendation for day T – 1
bankinter.returns
ALFREDO.recommendation.returns
Cumulative Return
0.05
0.00
−0.05
−0.10
−0.15
Daily Return
0.04
0.02
0.00
−0.04
Drawdown
0.00
−0.05
−0.10
−0.15
−0.20
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 27

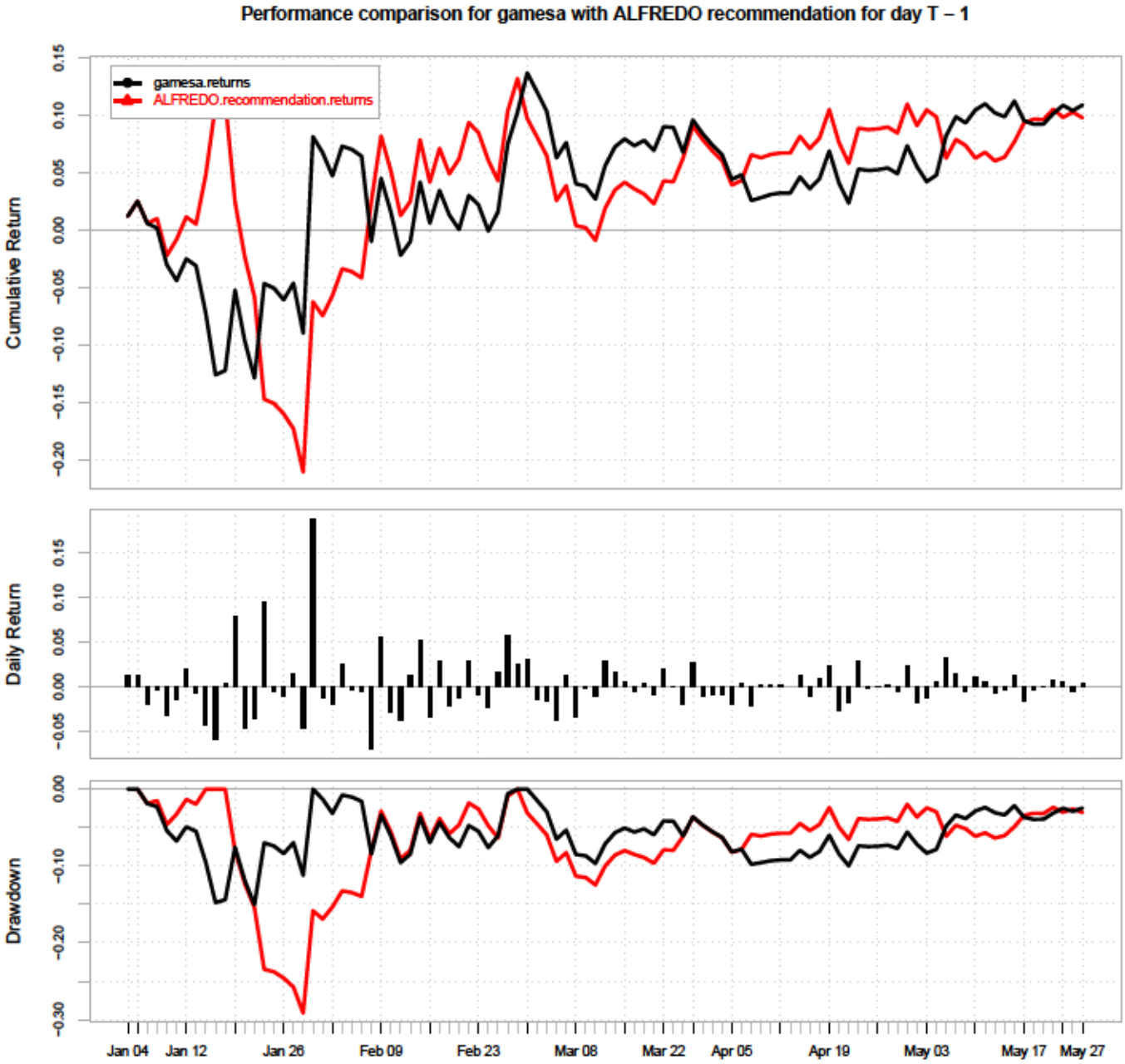

Performance comparison for bbva with ALFREDO recommendation for day T – 1
bbva.returns
ALFREDO.recommendation.returns
Cumulative Return
Daily Return
Drawdown
0.2
0.1
0.0
-0.1
-0.2
0.06
0.02
-0.02
-0.06
0.00
-0.05
-0.10
-0.15
-0.20
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31

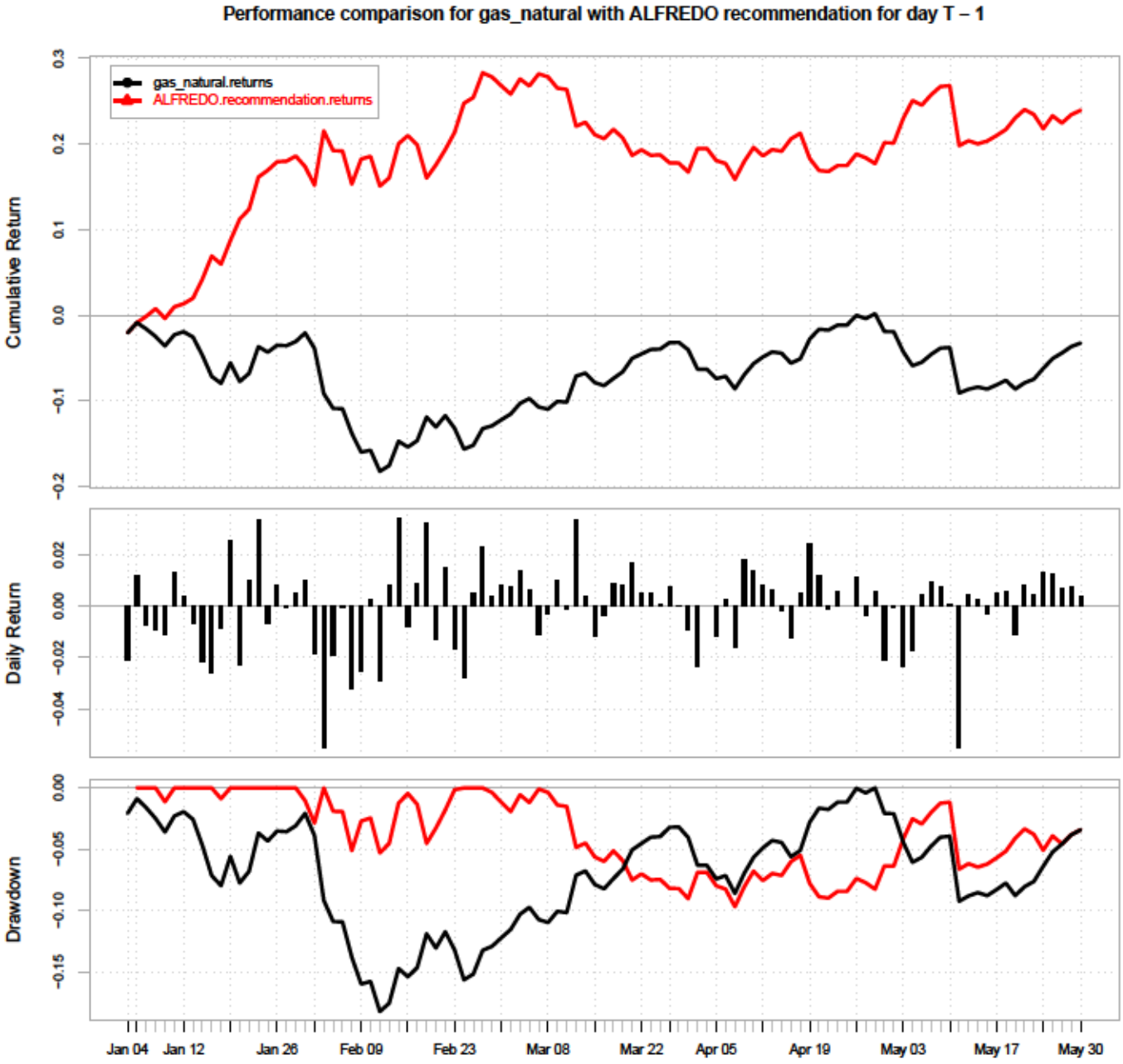

Performance comparison for cellnex with ALFREDO recommendation for day T – 1
cellnex.returns
ALFREDO.recommendation.returns
Cumulative Return
0.2
0.1
0.0
−0.1
−0.2
Daily Return
0.02
0.00
−0.02
−0.06
Drawdown
0.00
−0.05
−0.10
−0.15
−0.20
Jan 04
Jan 19
Feb 02
Feb 17
Mar 01
Mar 16
Mar 29
Apr 12
Apr 26
May 10
May 24

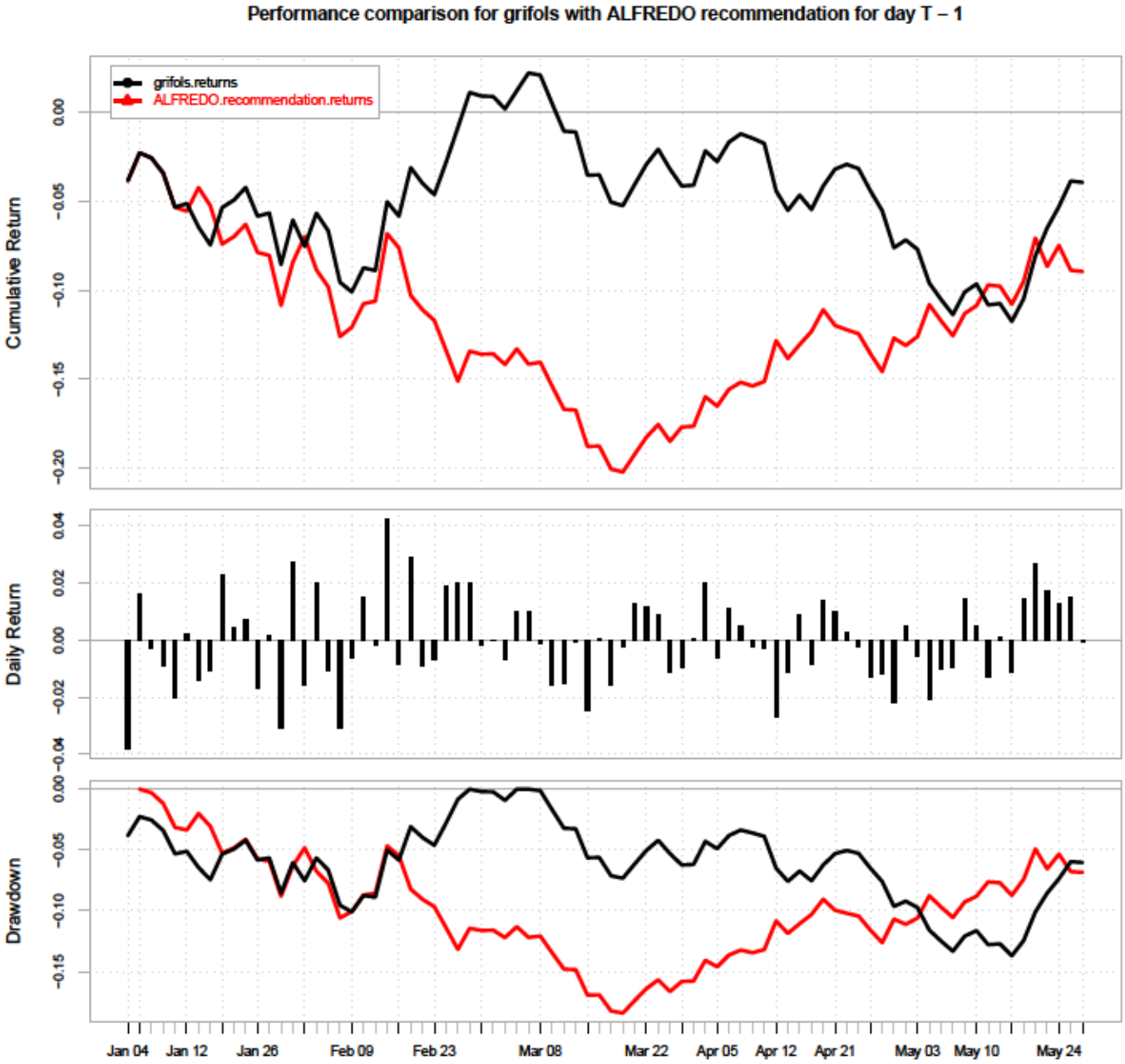

Performance comparison for enagas with ALFREDO recommendation for day T – 1
enagas.returns
ALFREDO.recommendation.returns
Cumulative Return
0.04
0.02
0.00
-0.02
-0.04
Weekly Return
0.02
0.01
0.00
-0.02
Drawdown
0.00
-0.02
-0.04
-0.06
-0.08
Jan 08
Jan 19
Feb 09
Feb 16
Mar 01
Mar 14
Mar 29
Apr 08
Apr 19
May 03
May 30

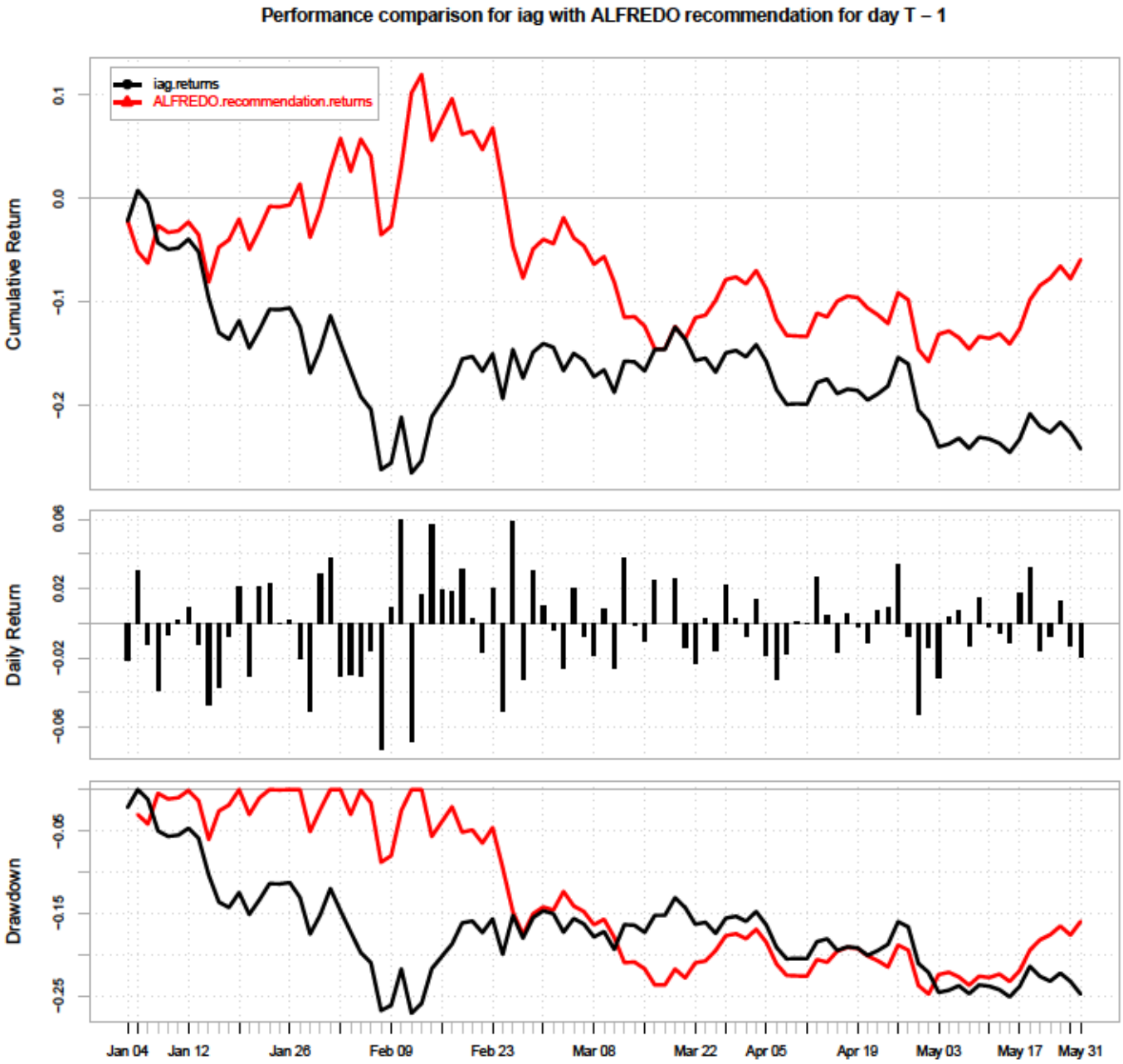

Performance comparison for endesa with ALFREDO recommendation for day T – 1
endesa.returns
ALFREDO.recommendation.returns
Cumulative Return
0.2
0.1
0.0
-0.1
Daily Return
0.03
0.01
-0.01
-0.03
Drawdown
-0.02
-0.06
-0.10
-0.14
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31

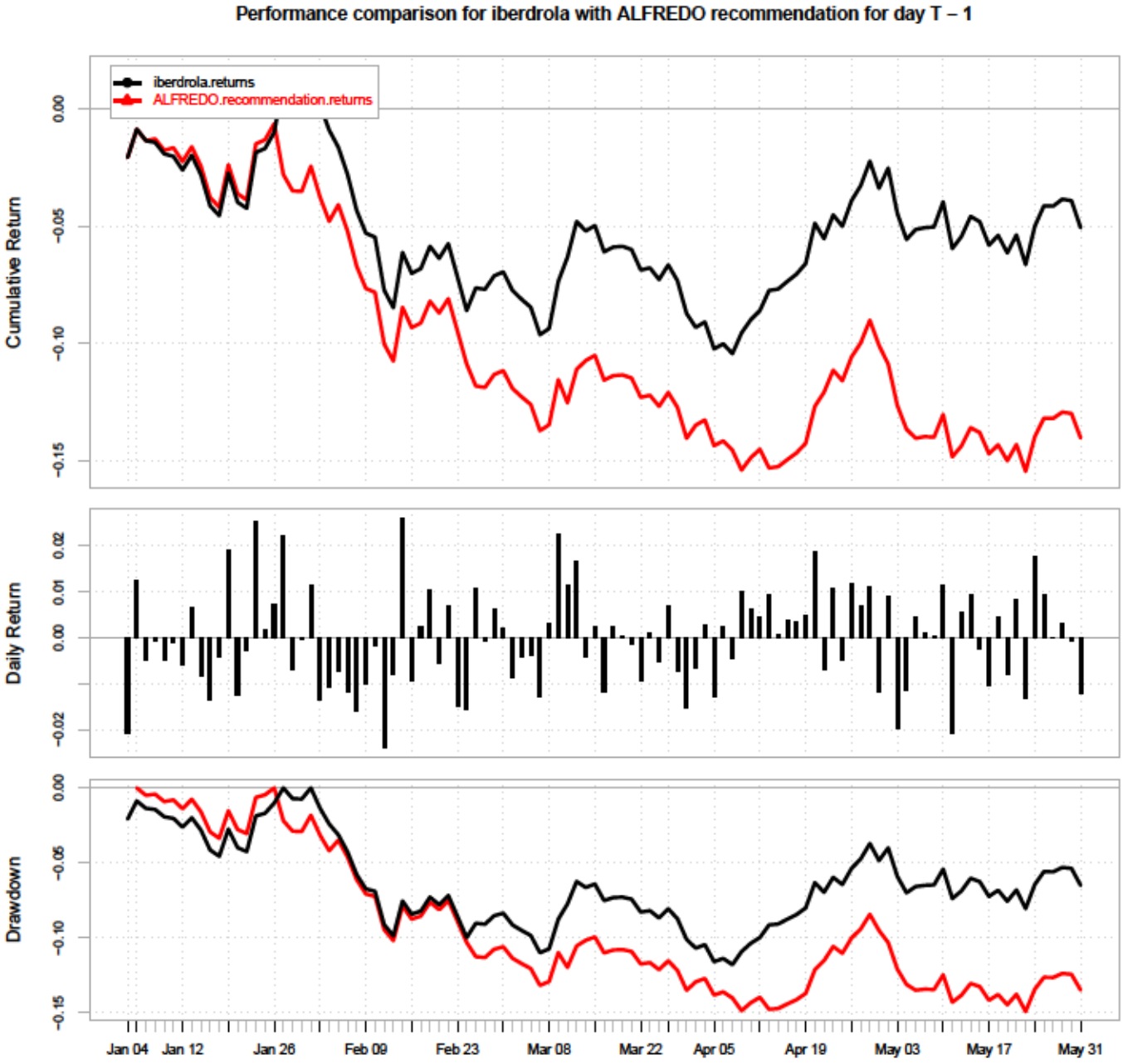

Performance comparison for ferrovial with ALFREDO recommendation for day T – 1
ferrovial.returns
ALFREDO.recommendation.returns
Cumulative Return
0.00
-0.05
-0.10
-0.15
-0.20
Daily Return
0.04
0.02
0.00
-0.02
-0.04
Drawdown
0.00
-0.05
-0.10
-0.15
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31

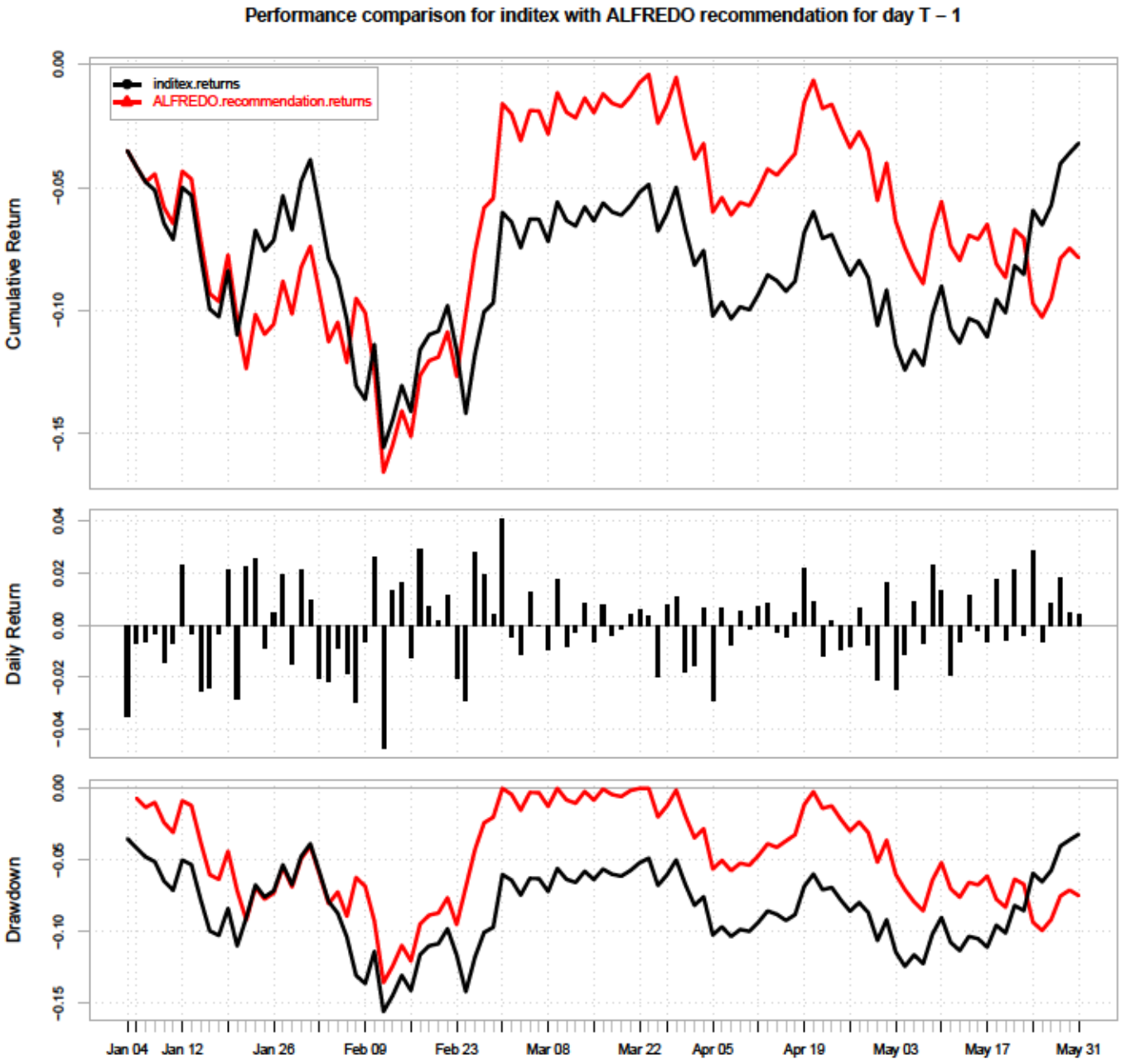

Performance comparison for gamesa with ALFREDO recommendation for day T – 1
gamesa.returns
ALFREDO.recommendation.returns
Cumulative Return
Daily Return
Drawdown
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.30
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 27

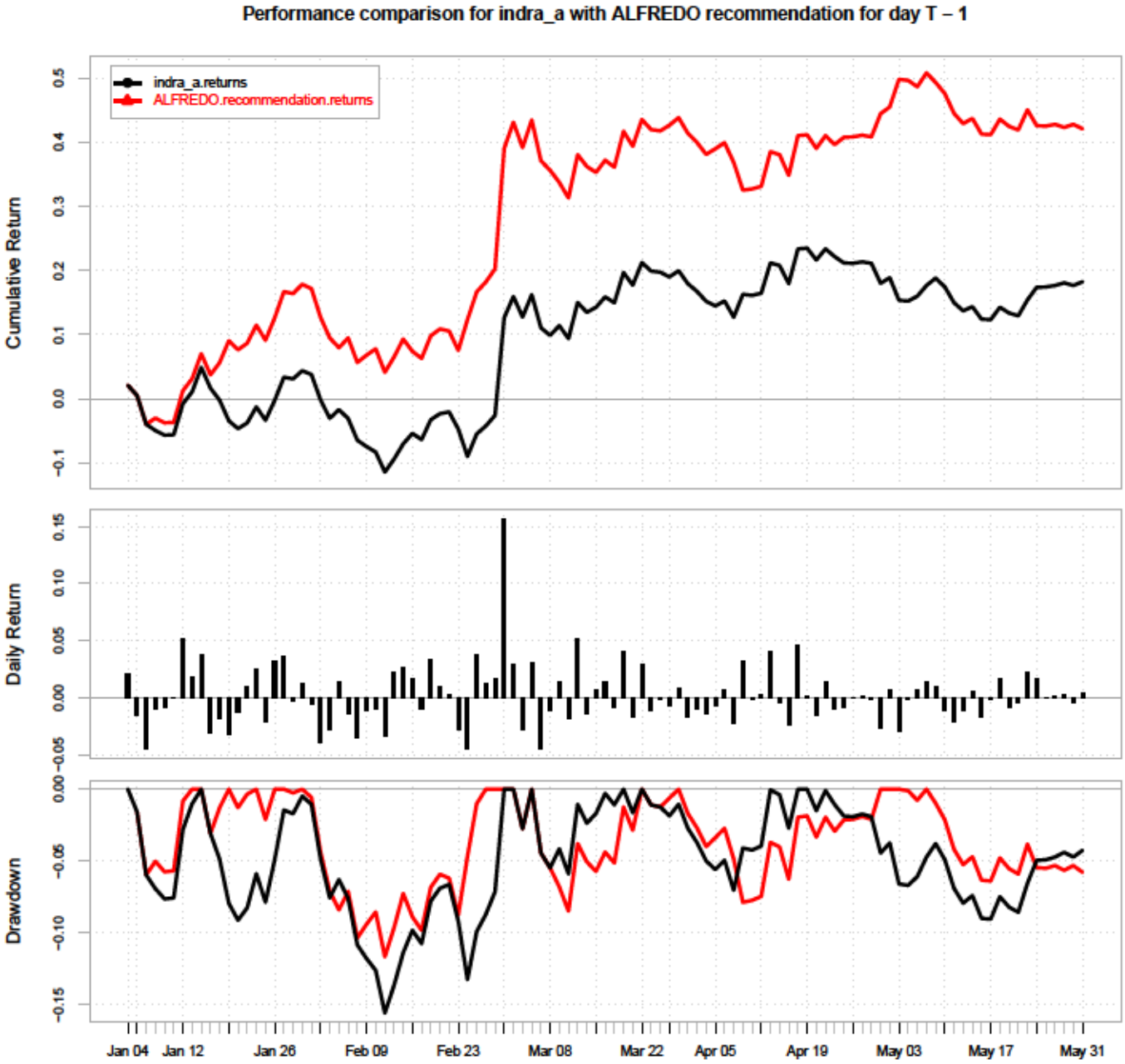

Performance comparison for gas_natural with ALFREDO recommendation for day T – 1
gas_natural.returns
ALFREDO.recommendation.returns
Cumulative Return
Daily Return
Drawdown
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 30

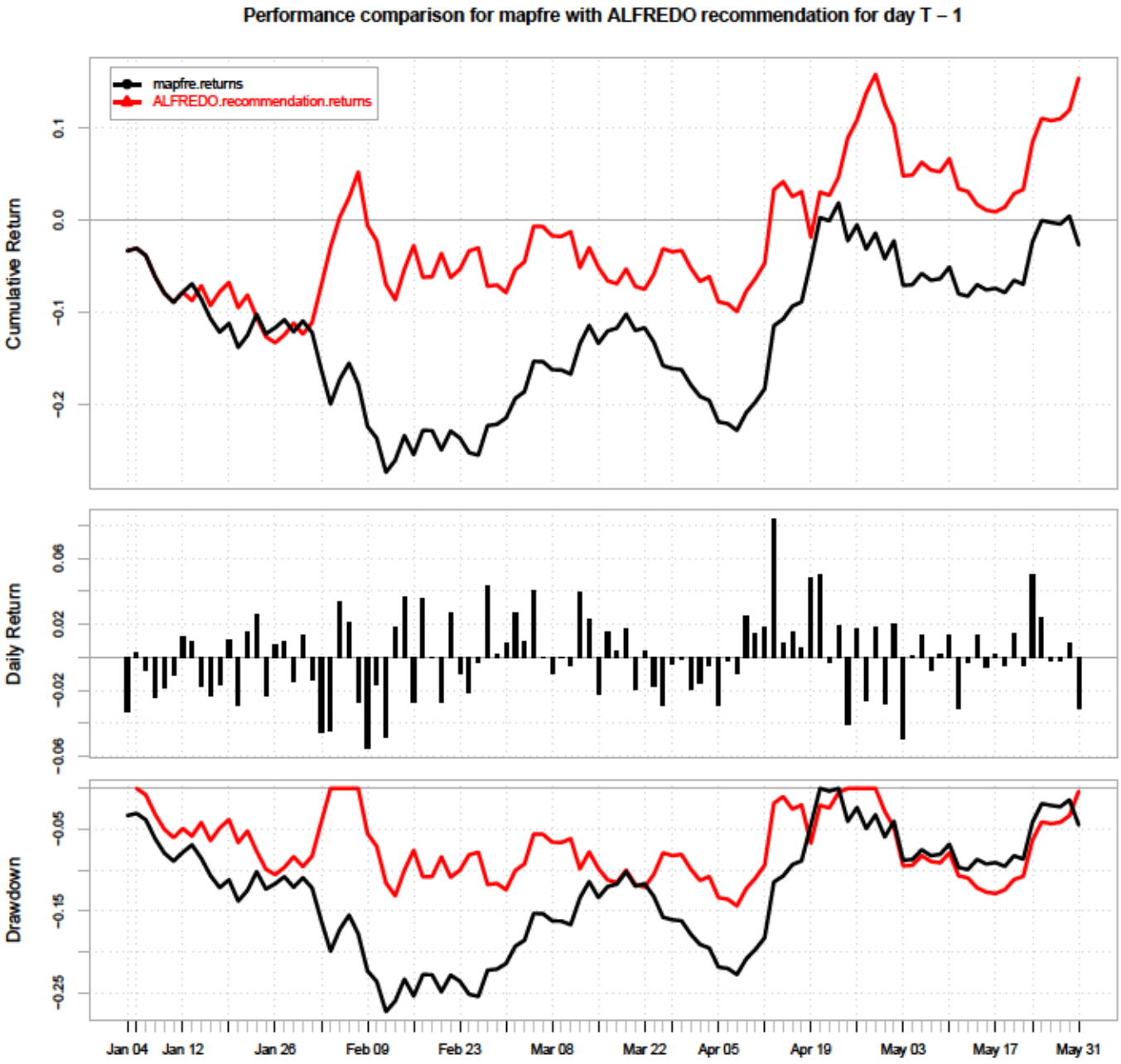

Performance comparison for grifols with ALFREDO recommendation for day T – 1
grifols.returns
ALFREDO.recommendation.returns
Cumulative Return
0.00
−0.05
−0.10
−0.15
−0.20
Daily Return
0.04
0.02
0.00
−0.02
−0.04
Drawdown
0.00
−0.05
−0.10
−0.15
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 12
Apr 21
May 03
May 10
May 24

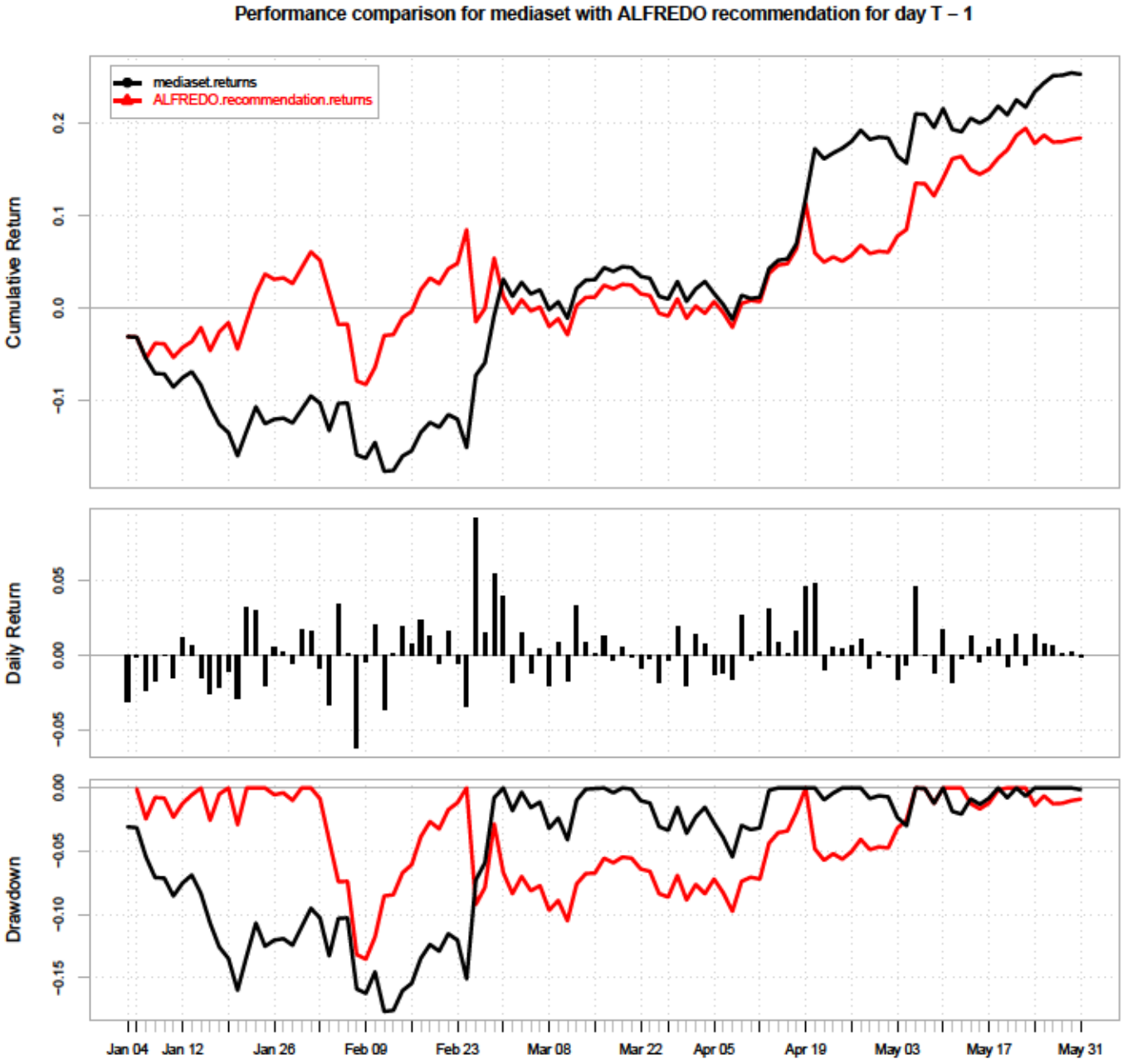

Performance comparison for iag with ALFREDO recommendation for day T – 1
iag.returns
ALFREDO.recommendation.returns
Cumulative Return
0.1
0.0
-0.1
-0.2
Daily Return
0.06
0.02
-0.02
-0.06
Drawdown
-0.05
-0.15
-0.25
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31

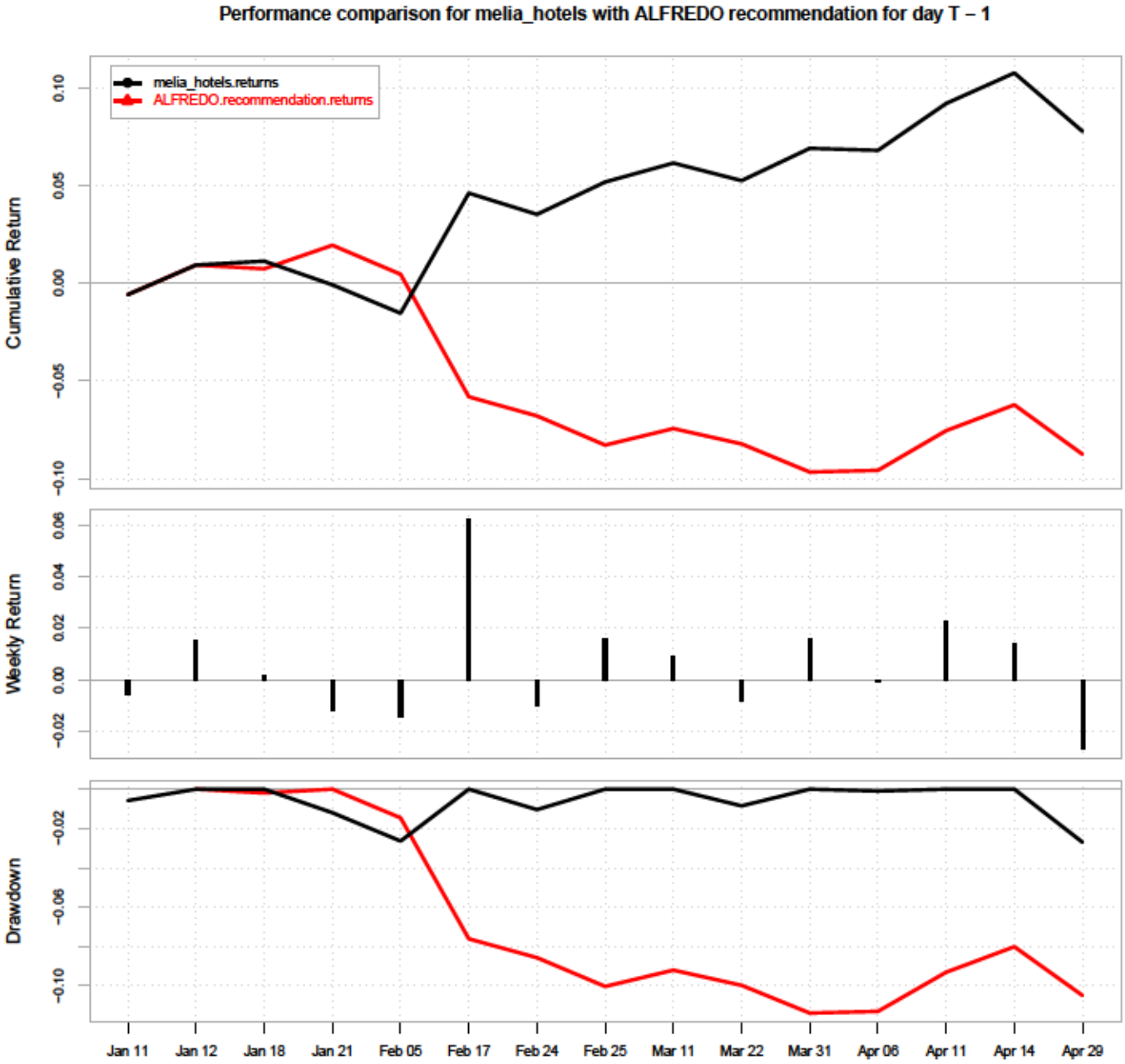

Performance comparison for iberdrola with ALFREDO recommendation for day T – 1
iberdrola.returns
ALFREDO.recommendation.returns
Cumulative Return
0.00
-0.05
-0.10
-0.15
Daily Return
0.02
0.01
0.00
-0.02
Drawdown
0.00
-0.05
-0.10
-0.15
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31

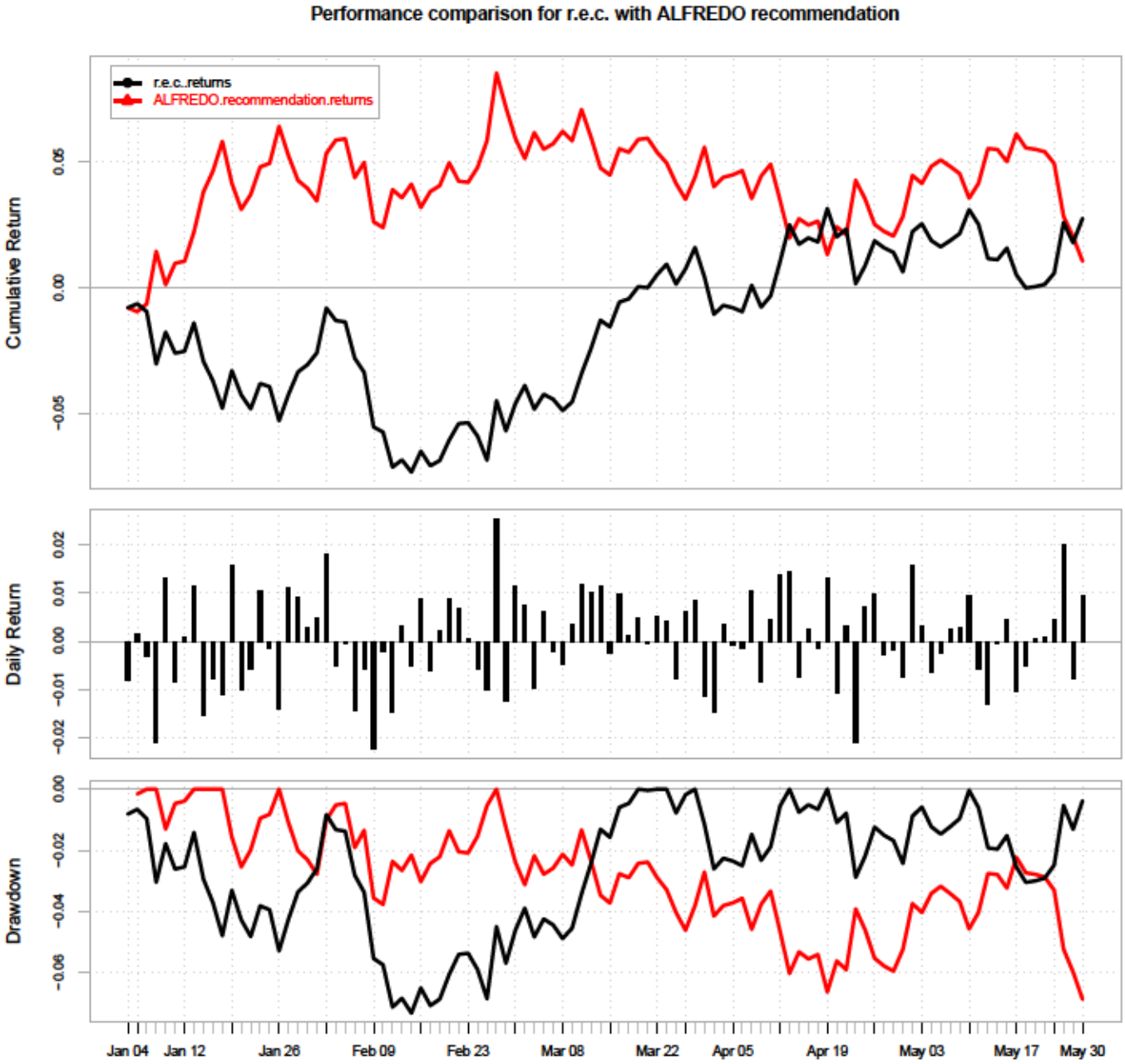

Performance comparison for inditex with ALFREDO recommendation for day T – 1
inditex.returns
ALFREDO.recommendation.returns
Cumulative Return
Daily Return
Drawdown
0.00
-0.05
-0.10
-0.15
0.04
0.02
0.00
-0.02
-0.04
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31

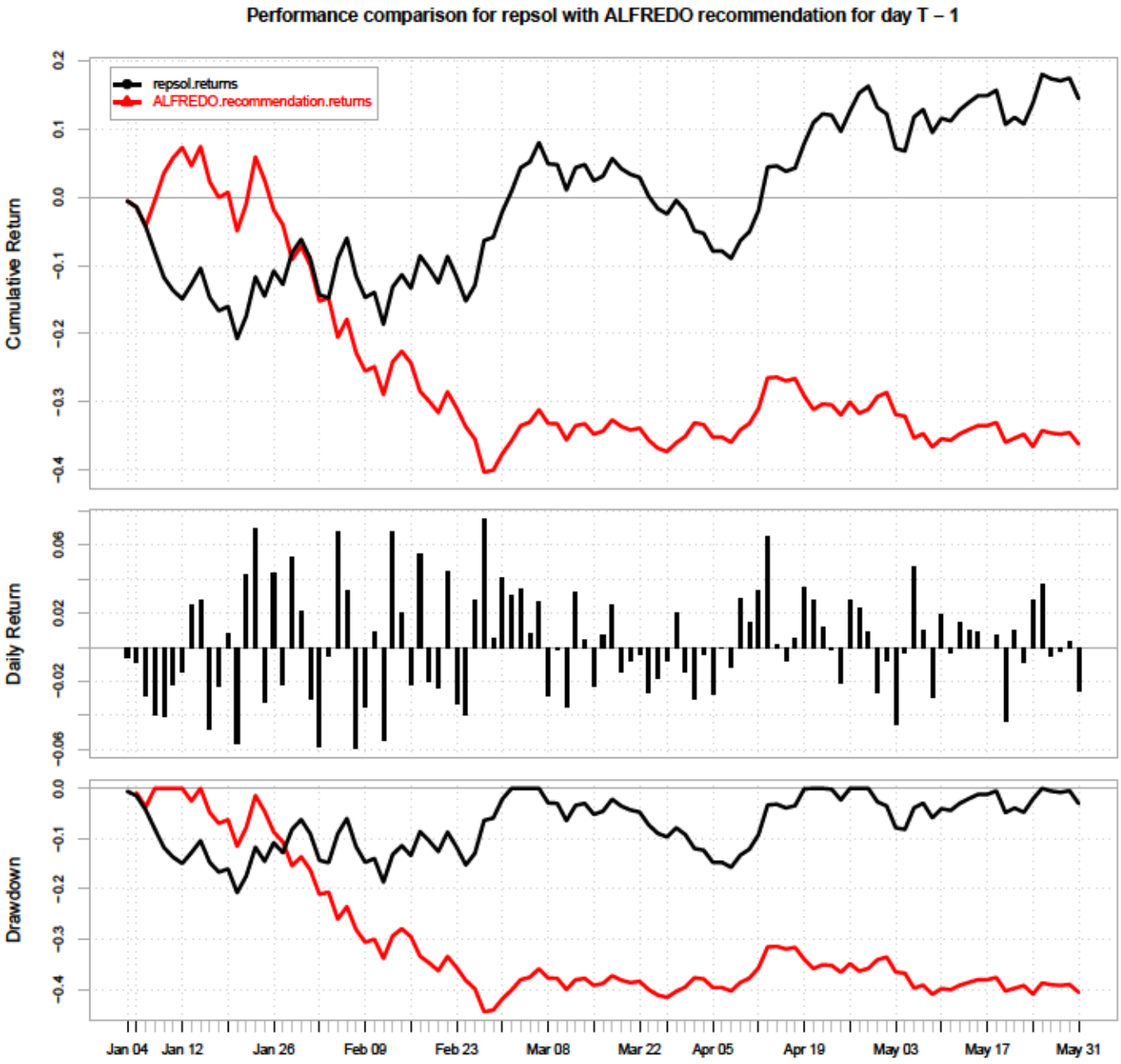

Performance comparison for indra_a with ALFREDO recommendation for day T – 1
indra_a.returns
ALFREDO.recommendation.returns
Cumulative Return
Daily Return
Drawdown
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31

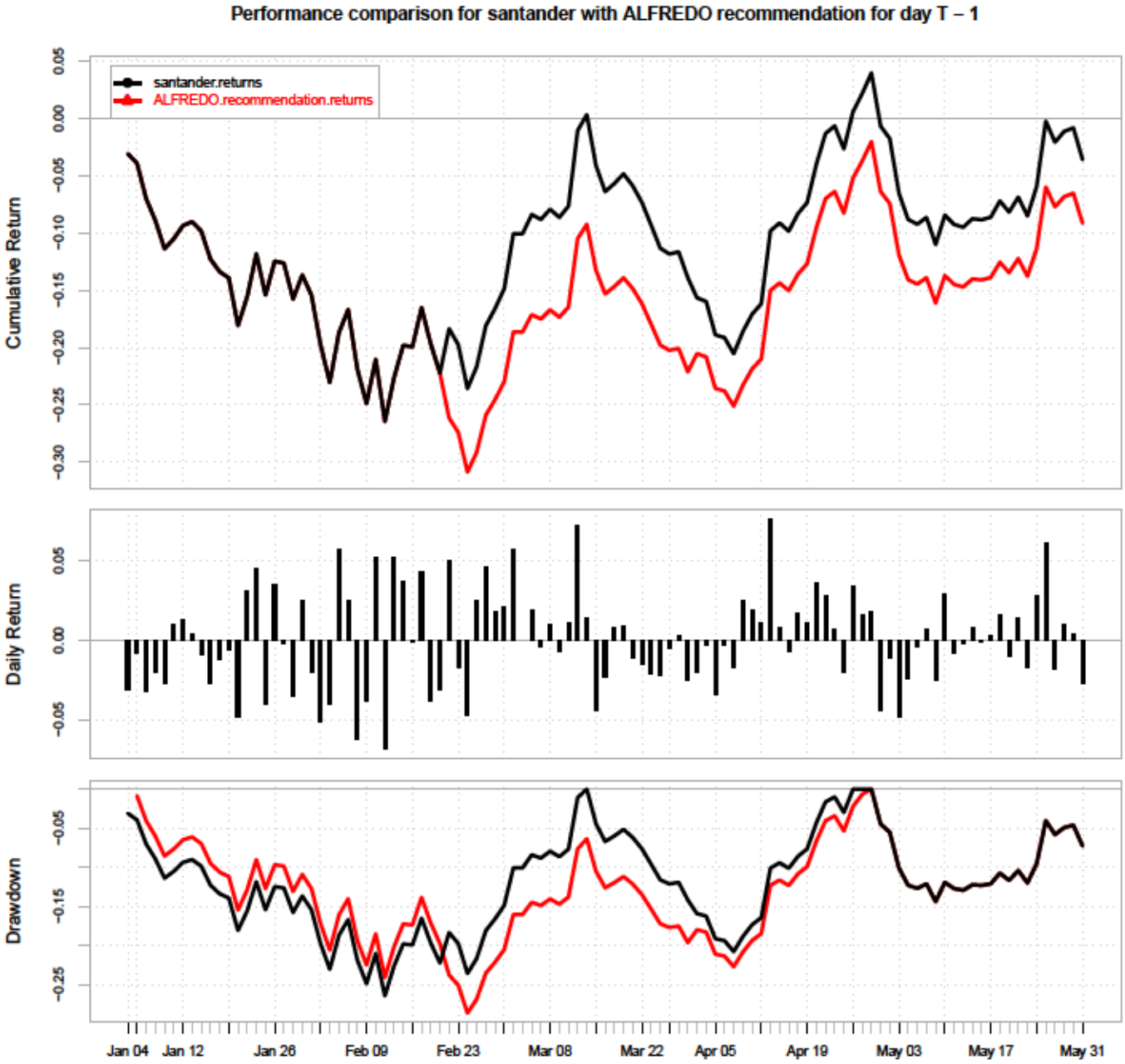

Performance comparison for mapfre with ALFREDO recommendation for day T – 1
mapfre.returns
ALFREDO.recommendation.returns
Cumulative Return
0.1
0.0
-0.1
-0.2
Daily Return
0.06
0.02
-0.02
-0.06
Drawdown
-0.05
-0.15
-0.25
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31

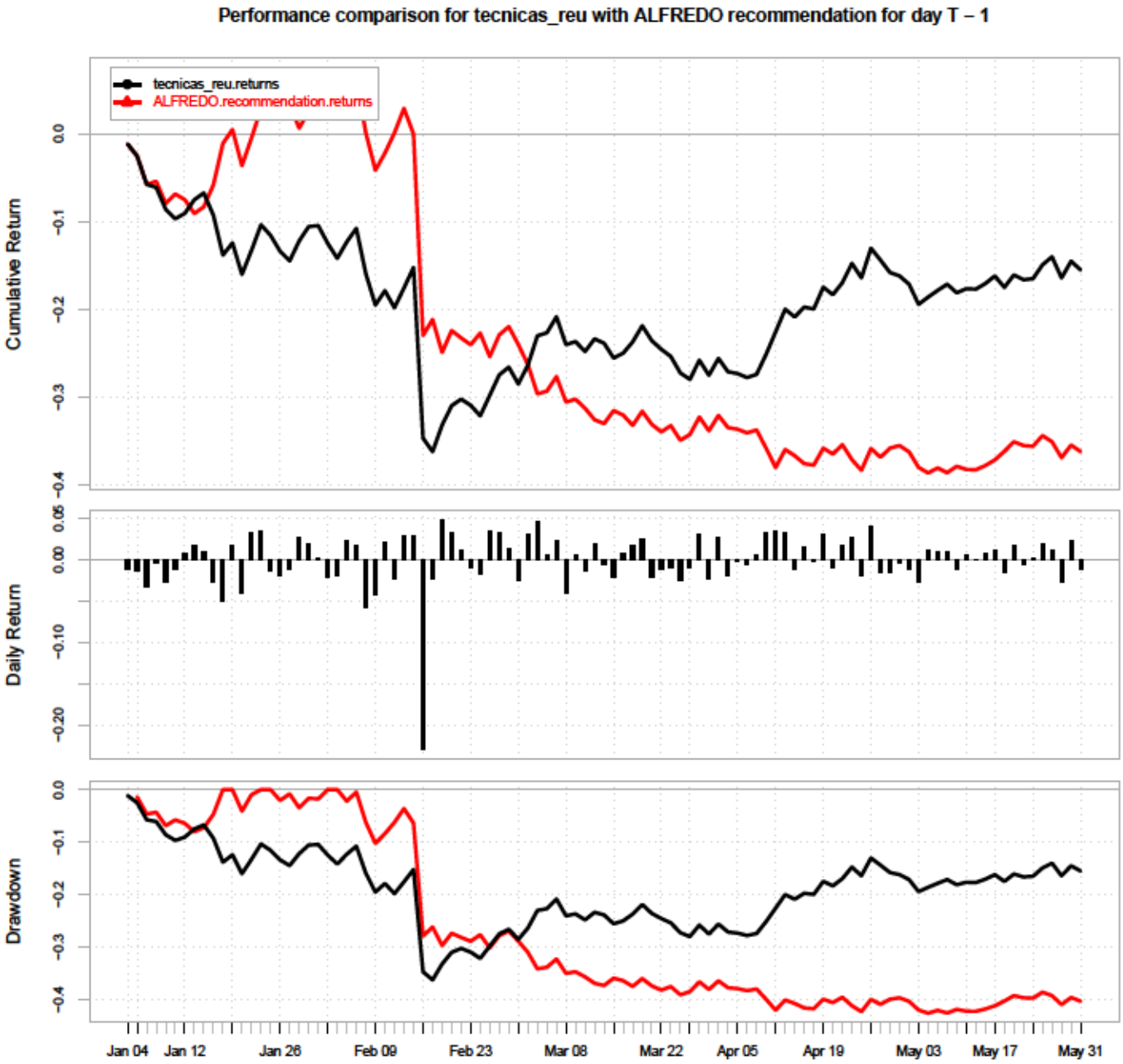

Performance comparison for mediaset with ALFREDO recommendation for day T – 1
mediaset.returns
ALFREDO.recommendation.returns
Cumulative Return
0.2
0.1
0.0
−0.1
Daily Return
0.05
0.00
−0.05
Drawdown
0.00
−0.05
−0.10
−0.15
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31

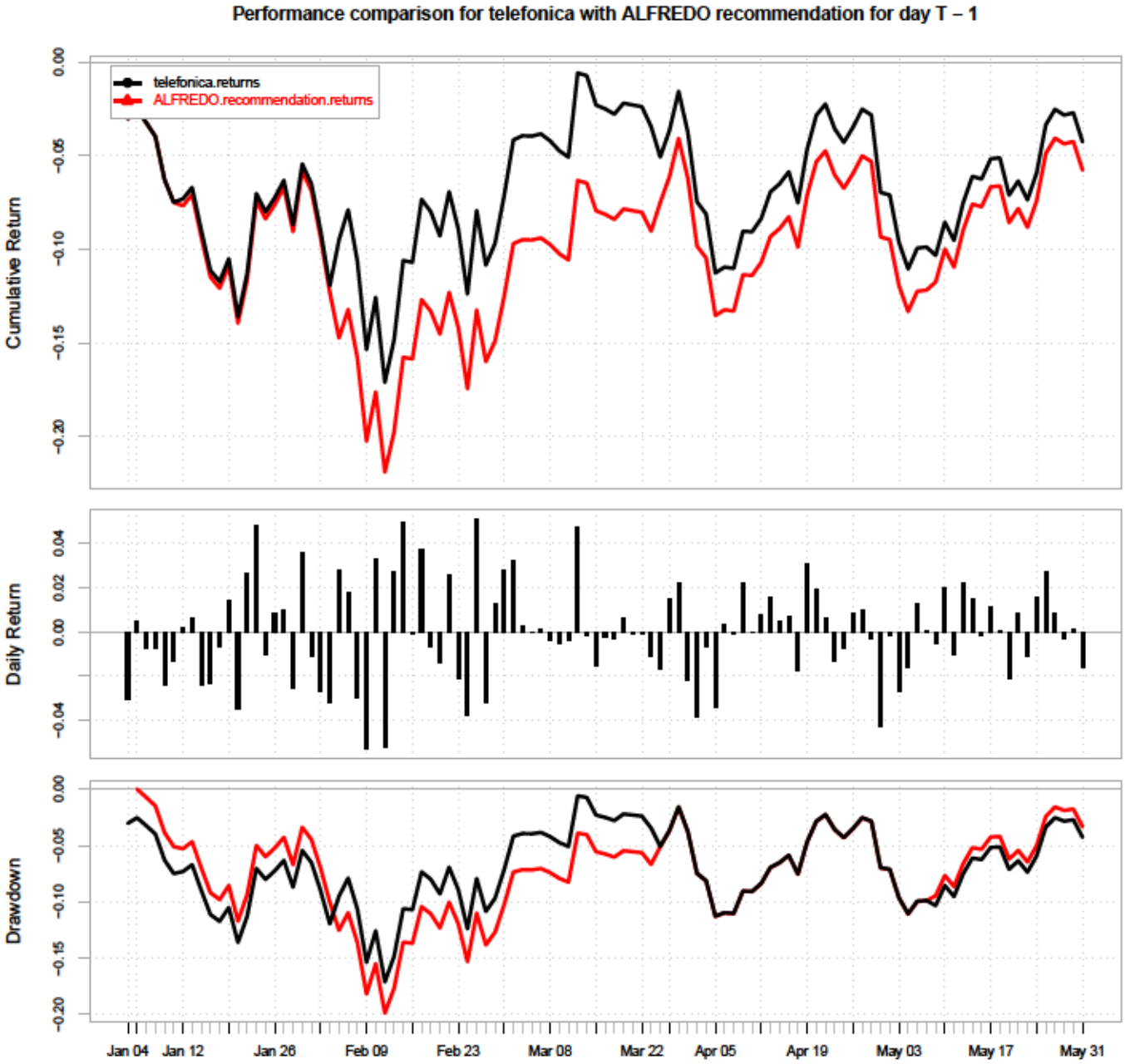

Performance comparison for melia_hotels with ALFREDO recommendation for day T – 1
melia_hotels.returns
ALFREDO.recommendation.returns
Cumulative Return
0.10
0.05
0.00
−0.05
−0.10
Weekly Return
0.06
0.04
0.02
0.00
−0.02
Drawdown
−0.02
−0.06
−0.10
Jan 11
Jan 12
Jan 18
Jan 21
Feb 05
Feb 17
Feb 24
Feb 25
Mar 11
Mar 22
Mar 31
Apr 06
Apr 11
Apr 14
Apr 29

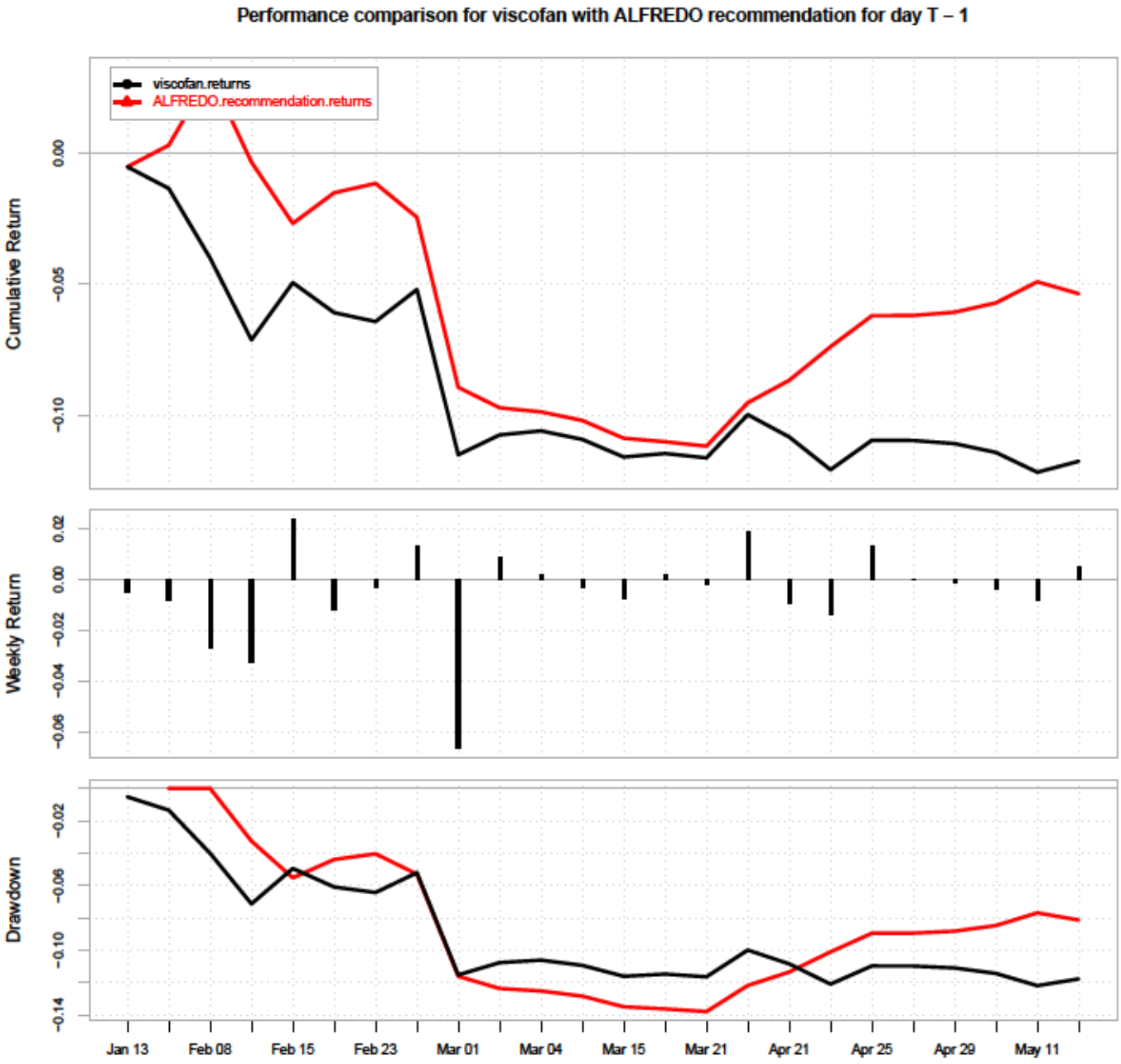

Performance comparison for r.e.c. with ALFREDO recommendation
r.e.c..returns
ALFREDO.recommendation.returns
Cumulative Return
0.05
0.00
−0.05
Daily Return
0.02
0.01
0.00
−0.01
−0.02
Drawdown
0.00
−0.02
−0.04
−0.06
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 30


Performance comparison for repsol with ALFREDO recommendation for day T – 1
repsol.returns
ALFREDO.recommendation.returns
Cumulative Return
Daily Return
Drawdown
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31


Performance comparison for santander with ALFREDO recommendation for day T – 1
santander.returns
ALFREDO.recommendation.returns
Cumulative Return
Daily Return
Drawdown
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31


Performance comparison for tecnicas_reu with ALFREDO recommendation for day T – 1
tecnicas_reu.returns
ALFREDO.recommendation.returns
Cumulative Return
Daily Return
Drawdown
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31


Performance comparison for telefonica with ALFREDO recommendation for day T – 1
telefonica.returns
ALFREDO.recommendation.returns
Cumulative Return
Daily Return
Drawdown
0.00
-0.05
-0.10
-0.15
-0.20
0.04
0.02
0.00
-0.04
Jan 04
Jan 12
Jan 26
Feb 09
Feb 23
Mar 08
Mar 22
Apr 05
Apr 19
May 03
May 17
May 31


Performance comparison for viscofan with ALFREDO recommendation for day T – 1
viscofan.returns
ALFREDO.recommendation.returns
Cumulative Return
0.00
-0.05
-0.10
Weekly Return
0.02
0.00
-0.02
-0.04
-0.06
Drawdown
-0.02
-0.06
-0.10
-0.14
Jan 13
Feb 08
Feb 15
Feb 23
Mar 01
Mar 04
Mar 15
Mar 21
Apr 21
Apr 25
Apr 29
May 11