UNIVERSITÀ DEGLI STUDI DI BERGAMO

Department of Management, Economics and Quantitative Methods

Applied Economics and Management (AEM)

Doctoral Thesis

Stock market portfolio strategies based on public news

Supervisor:
Prof. Sergio Ortobelli Lozza

Ph.D. Candidate:
Tommaso Adami

Referees

Associate professor Domenico DeGiovanni, Ph.D.
Università della Calabria

Associate professor Davide Radi, Ph.D.
Università di Pisa

Academic year

2019/2020

Dedicated to my family, my parents, and everyone else who made this possible.

Contents

1. Extended abstract

Public news is a relevant source of information describing events occurring in the real world. Despite information leakages and fake news readers are often willing to pay to gather information conveyed through the news. Fama's (see [Fam70]) Efficient Market Hypothesis (EMH) enumerates public news as one of the three pillars distinguishing the level of market efficiency.

In this work, we investigated the effects of micro news sentiment on the equity market. Therefore, according to several market theories of Behavioral Finance, we want to understand if public news is interesting for the equity market investor. The Adaptive Market Hypothesis (AMH) (see [Lo05]) is one of these theories and states that also if markets are efficient in general, inefficiency may affect them in particular periods. These inefficiencies may then leave open profit opportunities for the informed investor. For this reason, public news could have an important impact on investor's choices.

In the second chapter, we start by giving an introduction to Efficient Market Hypothesis, Behavioral Finance, and Adaptive Market Hypothesis. After the basic theoretical introduction, the chapter focuses on the literature review regarding available studies on the effect of the news on markets. The reported papers cover many relevant aspects regarding how to measure the news impact on the market returns. The main factors extracted from the news are:

- The news sentiment, that translates a news article into a numeric value regarding the expected news impact on the market.
- The news cumulated sentiment, that tries to reduce the noise present in the signal also due to the presence of fake news.
- The news relevance and novelty, that are scores assigned to news articles regarding the assets affected and the redundancy of the information.
- The news topics and categories, that are a high-level classification of news performed by most advanced news analysis tools, able to subdivide news articles into handcrafted subsets, characterized by different expected effects (sometimes also on different market sectors).

Our study is based on a database of news sentiment extracted from public news by Ravenpack and provided to us by Optirisk Systems Limited. The third chapter describes the market data and news sentiment database and reports statistics about the considered stocks. This research examines the news effects on the stock returns belonging to the EURO STOXX 50 market index at last once in the available considered period, between 1 January 2005 and 30 May 2018. The description of the data reports statistics about category groups, a high-level grouping of news events in a hierarchical taxonomy of financial-related news events developed by Ravenpack. The chapter also describes how daily cumulated sentiment indicators time series and news freshness weights are constructed from the original point-wide sentiment source.

The following two chapters describe the two models that constitute the main contribution of this thesis.

The fourth chapter presents news-based strategies for intraday open to close trading. The strategies are based on a naive beauty context model (see [Key36]) that takes into account only indicators generated by firm-specific public news sentiment or volume subdivided by category group. Three main patterns emerged from the analysis and characterize different category groups:

- Sell on news volumes.
- Buy according to news sentiment.
- Sell according to news sentiment.

The analysis has also shown that many portfolio strategies based on category groups are characterized by a reversal effect overnight, while few others by a continuation trend.

The fifth chapter presents a more complex model that tries to enhance a baseline model with the use of public news. Three baseline models are presented and the relative results compared. The baseline models rely on three different portfolio optimization criteria:

- Sharpe Ratio.
- Second-order Stochastic Dominance.
- Scaled Second-order Stochastic Dominance.

The stocks returns are regressed on market common trends to remove from the time series what is supposed to be market noise. The two Second-order Stochastic Dominance optimization criteria are based on enhanced indexation, where the optimal portfolio is supposed to be the best portfolio dominating the reference market index. For each baseline model, we regress the residuals of the market common trends separately on the principal components extracted from each category group. The aim of the study is to understand which category groups bring useful information for portfolio optimization. The discussion of the results is divided into two parts, the former reports profitability and turnover results for the three baseline strategies, while the latter reports the results for the different category groups and cumulation periods. In the last part of the chapter a series of possible further researches, that have appeared of interest, are proposed.

The last chapter report conclusions about the research.

Appendix A tries to describe the events happening in the market that could be conveyed through the news. At first, the events characterizing the value generation process and the highly complex firms' interconnection are described. The description then moves on to events concerning the firms' evaluation process, which is split into two sections: the former about fundamental indicators evaluation and the latter about perspectival views and ratings. The last section of the appendix reports events regarding the ownership structure of a firm and outlines a possible path of transmission from the news to the effects on the market.

Appendix B reports descriptive statistics about the news related indicators generated for the analysis.

2. Literature review

2.1. Introduction

In recent years, news has been widely used to predict market movements (and thus to improve asset allocation performance), as trends and volatility may find their drivers, or better, may be caused by real facts that could be reported in the news.

The markets, which according to classical literature should instantaneously incorporate every new information, as postulated by the Efficient Market Hypothesis, in some specific situations incorporate new information more slowly, allowing informed portfolio managers the possibility to exploit such inefficiencies. These inefficiencies are mainly supposed to be due to biased reactions to the news by traders with limited knowledge and bounded rationality. Due to such biases (i.e. overreaction, underreaction, or anchoring), traders tend to miss-evaluate the real effect of the reported events on asset prices. Such findings are coherent with the Adaptive Market Hypothesis, one of the possible explanations emerging to reconcile the Efficient Market Hypothesis and the more recent Behavioral Finance Theory. Under this hypothesis, markets are not constantly efficient, and the degree of efficiency depends on the composition of its participants and the available profit opportunities.

Market movement and market efficiency directly depend on investors' investment choices. Investors base their choices on available information that contributes to augment their beliefs. News and thus news sources are one of the main contributors to investors' available information and therefore these have been extensively investigated in the literature.

Many sources have been analyzed from different approaches, starting from the more official and authoritative, as official government board announcements, the main newspaper articles, and web scraping, to the vast amount of the real-time but unreliable and informal user-generated content of the social networks, like Facebook and Twitter.

A considerable number of tools have been developed to achieve better and better results in this kind of analysis. Financial industry firms, such as Thomson Reuters and RavenPack, have developed entire platforms capable of generating reports on news articles for each related stock showing the scores for main indicators that may reflect the effect on the market regarding relevance, novelty, coverage, and sentiment.

Different streams of research have been developed. The earlier ones statistically analyzed the presence and number of occurrences of particular words in the article.

More recent approaches try to disentangle more complex relationships from the sentences present in the articles, such as particularly relevant events.

The application of information extracted from news sources mainly regards market movements, for instance:

- Expected returns.
- Volatility.
- Liquidity.
- Volumes.
- Jumps.
- Value at risk estimates.
- Asset correlations.
- Financial community networks.
- Geopolitical risk indexes.
- Uncertainty levels in different periods.

News-related indicators can be inspected under different time frames. In regard to considered literature, authors focusing mostly on trading algorithms, such as machine learning, tend to analyze the effect on a short time frame, especially for trading purposes, often also on high frequency, while authors more interested in the theoretical aspects of the analysis tend to analyze the data on a longer time frame, where weekly or monthly effects can be exploited.

A relevant stream of research tries to analyze the news as already pre-processed information produced by externalized platforms such as Thomson Reuters News Analytics, Machine Readable News Platform (TRNA-MRN), or RavenPack services. Under this approach, for each stock, it is given a news feed containing numerical values describing the relevance, sentiment, novelty, and coverage from many sources for each article in the considered period. In this case, no other natural language processing technique is needed to extract information from the news. Those research papers show the potential and the limits of the news analysis where a text pre-processing platform is already present and then demonstrate the predictive capabilities of the news but also the limits, where for example in particular cases other information sources like the VIX could be shown to be more predictive than the pure sentiment.

Other branches try to analyze the news articles with increasingly more complex analysis techniques, starting from Bag of Words, going to Shallow Parsing, and event detection that is also the technique used in the previously mentioned TRNA-MRN platform.

2.2. The Efficient and the Adaptive Market Hypothesis

The most remarked and historically noteworthy theory regarding financial markets is the Fama's Efficient Market Hypothesis [Fam70]. Under EMH, all the information is immediately discounted by the markets, old information has no monetary value, and three forms of market efficiency are defined.

The first form of market efficiency is called *weak*. Under the *weak market efficiency* assumption prices and returns are considered to be old information. The second form of market efficiency is called *semi-strong*. Under *semi-strong market efficiency* assumption not only prices and returns are considered to be old information but also publicly available information is considered to be old information and should, therefore, be immediately discounted by financial markets. The third form of market efficiency is called *strong*. Under *strong market efficiency* assumption, also private information is considered to be old information, has no monetary value, and should, therefore, be immediately discounted by financial markets. Under those assumptions, investors cannot hope to beat the market, and no amount of analysis would help in generating abnormal returns.

The traditional theory, EMH, is based on the concept of *Homo Economicus*. Such a man should possess and base his choices on perfect rationality, perfect self-interest, and perfect information. He should also be motivated to optimize his marginal gains maximizing his Expected Utility Theory. Under the EMH, irrational traders are supposed to trade randomly and cancel each other, while arbitrageurs eliminate the remaining effects on prices.

A widely discussed alternative to the EMH is Behavioral Finance introduced by A. Tversky, D. Kahneman, and R. Thaler, whose debate spans over 40 years. However, as reported by [WKM12], also at the end of this debate, Behavioral Finance had difficulty taking place, and EMH remains a reference model.

Behavioral Finance ([Cop15]) is a branch of Social Psychology in which investors are supposed to be *Normal* rather than rational. A *Normal* investor does not possess the same qualities of the *Homo Economicus* while he posses the weaker specular qualities defined by Behavioral Finance. Instead of perfect rationality, he has bounded rationality, instead of perfect self-interest, he has agency problems, and instead of complete information, he has limited information. Due to biases in his reasonings, instead of optimizing his marginal gains, he assigns values to gain and losses. The investment process has been described by Behavioral Finance using a set of specular biased frameworks composed by the Prospect Theory, the Behavioral Portfolio Theory, and the Behavioral Asset Pricing Model. In Behavioral Finance, investors are then supposed to be biased irrational traders who have herding behaviour and arbitrage does not always take place because it is risky and therefore limited to profitable situations.

Fama himself criticized the strong market efficiency assumption, saying that even if not realistic it could be useful as a theoretical reference point. Fama also noted that uncertainty concerning intrinsic stock value generates noise trading, which in turn could produce bubbles. However, according to Fama, sophisticated traders will burst the bubbles before they even have a chance to develop. De Long et al. (1990) instead criticized the EMH, saying that it ignores the systematic risk of noise traders that could lead to significant divergence and that arbitragers may have limited resources needing to liquidate their positions before prices reverts. This fact could also explain why examining pseudo-signals followed by noise traders could be profitable.

Grossman and Stiglitz also criticized EMH in 1980, bringing as an argument the impossibility of informationally efficient markets. They argued that under the assumption of perfectly efficient markets, investors do not have any incentive to acquire costly information if markets are not inefficient, and thus, there are no profit-making opportunities available.

The idea of markets driven by changing dynamic forces had already been postulated, as reported by [Egi14], by Minsky a post-Keynesian economist. [Min92] identified three kinds of economic units: hedge, speculative, and Ponzi, asserted that if hedge finance dominates, the economy is equilibrium-seeking, while in the other cases is deviation-amplifying. He also suggested that under periods of growth, economies tend to move away from a hedge structure.

The Adaptive Market Hypothesis, AMH ([Lo05]), is one of the possible theories that emerged between the others and that may reconcile the Efficient Market Hypothesis and Behavioral Finance. The AMH is a market theory that allows the co-existence of EMH and Behavioral Finance Models under which market efficiency is related to environmental factors characterizing market ecologies such as the number of competitors in the market, the magnitude of profit opportunities available, and the composition and adaptability of the market participants. Typical market players in financial markets can be stereotyped in well-known classes: pension fund, retail investors, hedge managers, hedgers, speculators, arbitragers, and market makers. An example of different markets characterized by different market environments could be well represented by the contraposition between the highly efficient market of the ten years US Treasury Bonds and the less efficient and liquid market of the Italian Renaissance Oil Paintings.

The efficiency of different markets has been widely studied in recent years. The advent of the Adaptive Market Hypothesis has shifted researchers from an all-or-nothing approach to one where the degree of efficiency is measured as a dynamic factor. The degree of efficiency has been measured using several linear and non-linear statistical tests. The main aim of these tests is to falsify the Random Walk Hypothesis or the Martingale Difference Hypothesis. If the returns are not independent, or the expected value of the returns conditioned by the past returns is not a constant, that means that to some extent markets are predictable and this violates the weak-form efficient market assumptions.

As a consequence of the emergence of AMH, researchers have analyzed the degree of market efficiency. Many different markets have been tested, and empirical evidence supporting the AMH has been found. [UM16] tested the efficiency of S&P500, FTSE100, NIKKEI225 and EURO STOXX 50 between 1990 and 2014. The authors evaluated time-varying return predictability using linear and non-linear tests with a two years rolling window. They also tried to find relationships between predictability and market conditions. The results have been found consistent with the AMH. The return predictability fluctuates over time in each market alternating periods of predictability and non-predictability. The markets evolve differently over time, and the predictabilities are not very correlated. The S&P500 has been identified as the most efficient, while the EURO STOXX 50 as the least. The results suggest that each market interacts differently with the considered market conditions, such as bull, normal, or bear market, up or down periods, and high or low volatility.

[CDK12] analyzed the US, UK, and Japanese markets using very long-run data applying linear and non-linear tests with a five years window. After running each test, the authors have also classified markets in categories, such as "inefficient", "AMH" or "switch to efficiency". The results are consistent with the AMH, with the linear tests showing periods of independence and dependence, and non-linear tests showing strong dependence for every subsample with time-varying magnitude. The research found very little evidence of a switch to efficiency. The non-linear tests efficiency degree has not remarkably declined over time.

[KSL11] studied the predictability of the Dow Jones Industrial Average index from 1900 to 2009 applying linear and non-linear tests with a two years window for daily returns and a five years window for weekly returns. The results have been found consistent with the AMH, and the predictability driven by changing market conditions. A high degree of uncertainty and high volatility have been associated with returns predictability, as also political and economic crises with moderate uncertainty, while market crashes have not. On the other hand, smaller predictability than in regular periods has been found correlated to market bubbles. Also, economic fundamentals, and in detail interest rates and inflation, have been discovered to be correlated with predictability. Inflation makes forecasting difficult and thus lower predictability, while the market results to be more predictable with higher interest rates. The year 1980, has been found to represent a breaking point in returns predictability results. Before 1980, returns have a higher degree of predictability, while the market efficiency has improved since then. The authors suggest that the improved efficiency may be due to two factors. The former is that in the 1960s and 1970s the market implemented a series of innovations that seems to have gradually taken place, for instance, the automation of trading, new regulations, and the establishment of a futures market. The latter is that the breaking point is also consistent with *Great Moderation* a decline in the volatility of macroeconomic fundamentals in the US.

[UM14] analyzed four calendar anomalies in the Dow Jones Industrial Average index from 1900 to 2013 and found that these support the AMH. The calendar performances of the Monday, January, Turn-of-the-Month, and Halloween effects have been found time-varying and dependent on the considered market conditions.

[HM13] studied the predictability of emerging markets from 1995 to 2011 using a four years rolling window. The authors found evidence of long-term memory for volatility and weak evidence in the case of returns. They divided the markets into two groups *Advanced* and *Secondary*. Their findings constitute evidence against weak-form EMH, are consistent with AMH, and suggest that *advanced* emerging markets are subject to less memory persistence. The results also conflict with the Grossman and Stiglitz argument that markets tend towards efficiency over time, because the levels of the 22 considered emerging market widely fluctuates, but some have an uptrend, while others a down one.

[CDK12] tested the exchange-rate return predictability of five independently floating currencies relative to the US dollar from 1975 to 2009 using a two years window. The results show that exchange rates are generally unpredictable, but episodes of predictability exist, and consistently with the AMH they are associated with changing market conditions, that may happen during events like central bank interventions and financial crises.

2.3. The news

2.3.1. The news impact

Many authors studied how news impacts market prices. As shown by many authors, a well-known difference in the way news impacts the markets is that news effects are different among the main geopolitical areas, that could be individuated as United States, Europe, Australia, Japan, and Asia ex-Japan.

The impact of the news was also found different among different industries, as among different sectors. Consistently with the Adaptive Market Hypothesis, these can be characterized by different types of investors, and therefore have different reactions to the same type of news. [SW11] studied the news effect separately for each industry sector, using the first 2 digits of the Standard Industrial Classification (SIC). As reported in [RLJW18], studies from Wolfe Research highlighted that Technology, Media and Telecommunications stocks are particularly sensitive to news coverage and, as reported in [LRW+18], that news sentiment data have additional insights for banking stocks. [SYY17] reported different reactions after pre-IPO announcements for electronics and non-electronics industries. [HGC15] found overlapping but different results between the US and Europe and between small and large/mid-capitalization companies.

[BFKR13] analyzed the market reaction at different levels of complexity for the information reported in the news. The authors showed that investors tend to overreact to simple news, while, probably due to the incapacity to fully and correctly evaluate the effect of the news, tend to underreact to complex news.

Another critical factor to understand the news impact is the difference between attention-grabbing news and low-attention news. An interesting study of the attention-grabbing level on news was conducted by [BKLB14], who were able to analyze Yahoo Finance Web Browsing activity data. Researchers found that web browsing activity can anticipate stock trading volumes by two or three days. The results achieved better performances with hourly rather than daily data. Also [RBB$^+$14] and [RC15] found that intraday scale browsing activity is correlated to market movements and helps to predict future movements. Unfortunately, such data are not public and freely available, and have to be requested directly to Yahoo Inc.. [MCA$^+$13], instead, analyzed Wikipedia usage patterns to discover traces of attempts to gather information on stocks before trading decisions were taken and found that changes in Wikipedia company page views volumes may contain early signals of stock market moves and could allow gaining new insight into the information-gathering stage of decision making. A similar analysis was carried out by [Kri13], who analyzed search engines query volumes from Google Trends and found that stock popularity as measured by search queries is correlated with stock riskiness. Following this approach, the author found that a portfolio in which popular stocks are penalized while less popular are brought forward dominates the benchmark and an equally weighted portfolio. [HS16] found underreaction for low attention-grabbing news and that positive news is less attention-grabbing, possibly due to an investor's bias. These announcements may be incorporated into prices around the next earnings announcement. [SW11] also reports that the low attention-grabbing effect for positive news could also be related to the fact that firms allow positive news to leak more easily. [BO08] measured attention using social media activity, and found that high levels of attention are associated with greater sensitivity of earnings-announcement returns to earning surprises, with a stronger effect for firms beating analysts' forecasts, while only firms with low levels of attention are associated with significant post-earnings-announcement drift. The authors discovered that individual investors are more likely to overweight attention-grabbing stocks, irrespective of any new information, and this tends to lead to a quick reversal of price movements, while price movements driven by new information tend to persist in the long term. In [ADAWLS18], the authors found that social media activity - measured on Twitter and StockTwits - has a significant impact on liquidity at the intraday level, with negative sentiment having a much larger effect on liquidity measures than positive sentiment. The analysis also showed that, at the intraday level, peak social-media sentiment corresponds to the end of momentum and a return to mean reversion.

A different approach was used by [GIS17], who classified social media news from Twitter (tweets) as local or non-local. The author considered the tweets posted from the neighbourhood of a firm's headquarter as local news, and found that non-local tweets have negative return predictability, while local ones are more relevant for international markets.

[YMAVS15] state that news is an event that moves the market in a small or in a big way, and a novel concept is introduced: a derived measure of news impact, able to take into account the news flow and the time decay effect, cumulating the news discounted by an exponential decay factor based on news seniority.

In [MdBBY15], the authors found an indirect relationship between news and market microstructure (i.e. liquidity, bid-ask spread, trade size, and market depth) through the cross dependencies of volumes and volatility. If information comes as a scheduled announcement, market participants anticipate it, formulating conditional plans on the contents, could react quickly and liquidity is maintained. Instead, if unanticipated news is revealed to the market, financial market participants need some time to formulate appropriate actions. During the period of contemplation, traders are unwilling to trade, and liquidity dries up.

[HGC15] report that market response is consistent with sentiment across a ten-day period particularly for negative news events and small-capitalization stocks. Authors also mention that impact of news on stock prices is not always straightforward and three possible scenarios have to be expected: prices find equilibrium quickly with a fast signal decay, prices initially underreact, leading to continuation effect, or prices initially overshoot, leading to reversal effect.

[BFKR13] analyzed the difference between news and no news days and found that reversal happened on no news or unidentified news days, conditional on extreme moves, while identified news days are characterized by momentum continuation. The authors report that a strong contemporaneous response of media coverage on identified news days, but not in unidentified news days, could be interpreted as those days are days on which price-relevant information arrives. The authors also note that contemporaneous price response to identified news days is unlikely to be due to irrational overreaction to news coverage, but that could suggest that price response is insufficiently strong for many of the event types.

[MMD08] report that traditional multifactor risk models fail to update quickly as new information becomes available. The authors updated an existing model that was incorporating option-implied volatility to include also changing market sentiment. The changing sentiment was calculated as the 7 days cumulated daily 15 minutes variance of news sentiment score.

[KD12] found that consistently with the Mixture of Distribution Hypothesis (see [DLFM17]), the variance of returns is proportional to the rate of information arrival on the market. The authors argued that volatility clustering reflects the serial correlation of information arrival frequencies.

[DRT12] analyzed the asymmetry of volatility under different market regimes. The authors found that asymmetric stronger volatility for down-markets is most likely driven by the overreaction of private investors to bad news. The authors also highlighted that the effect is stronger in more developed markets, probably due to the presence of more private investors.

2.3.2. The news sentiment

Every news article is composed of words, phrases, and concepts, every one of which can be associated with a particular sentiment. Sentiment scores of the whole article can be evaluated using statistical techniques. The sentiment, usually positive, negative, or neutral, can be relative to the common sense associated with words when it is evaluated using a standard sentiment dictionary, more technical when evaluated using a domain-specific sentiment dictionary, or even task-specific when for instance it is calculated as the expected effect on some other factor. An example of a task-specific sentiment can be the NIP sentiment calculated by RavenPack representing the impact on stock price volatility in the two hours after the announcement.

Since earlier studies, one of the most evident properties of the sentiment effects, studied by many authors, has been that the effects are characterized by a strong asymmetry between positive and negative sentiment. Another important result, highlighted by many authors, is that the cumulated effect of the news sentiment is a significant predictor for market movements and that a longer cumulative period corresponds to a longer predictive effect. [HS16] and [ZHCB16] showed that negative news has a more prominent and longer-lasting effect than positive news. In detail, [HS16] found that daily cumulated news sentiment is correlated to market movements in the following 1 or 2 days, while weekly cumulated sentiment can predict returns up to a quarter, or 13 weeks, in case of negative sentiment, and up to one week in case of positive sentiment. [ZHCB16] also suppose that the movement relative to the underreacted part of the positive news could be incorporated into market prices near next-earning-announcement, as this kind of information will emerge in this type of announcement.

[GKG⁺14], [GKG⁺15] and [KGU⁺15] report that on intraday scale the markets seem to be informationally efficient in respect to news sentiment, while on a longer scale feedback is present.

[HS16] compared the returns distributions of stocks without news, with news, and with neutral news, and found that stocks with neutral news have better performances than stocks without news, contradicting the proverbial phrase “No news is good news”.

[HLGC⁺18b] compared different ways to aggregate daily news sentiment and introduced *Sum Excess Sentiment Indicator* (SESI). SESI is a simple way to take into account sentiment and volume simultaneously. It is constructed by daily cumulating the fresh and relevant news events sentiment after subtracting the cross-sectional daily sentiment bias per stock universe. The indicator considers only news events highly relevant for the relative stock with no similar news in the last 90 days. The cross-sectional sentiment bias is a daily average of news events sentiment using a slightly different weighting for news event relevance and freshness. The authors report that the indicator also naturally includes *Buzz Effect*, i.e. ranking by sentiment strength. Trading strategies obtained by SESI were compared with these obtained by an average sentiment indicator. The results have shown that SESI overperformed the average sentiment indicator for both annualized returns and information ratio. The authors also inspected through a simulation the news strategy scalability. The results showed a decrease in information ratio and annualized returns between 100 million dollars and approaching 10 billion dollars.

2.3.3. The news correlation with returns

[ES17] and [ES18b] studied the correlation between the sovereign bond yield spread and macroeconomic news sentiment. She found that yield spread, spread change, and spread volatility are correlated to daily macroeconomic news volume and sentiment. She also found that a high percentage of the rolling correlation with a 250 days window is significant and thereafter she proposes to monitor correlation changes to recognize changing market conditions. She reports that a change in the sign of correlation points to changes in market environments; i.e. positive and negative sentiment series correlation changes; that the strength is based on the business cycle. The change of sign of the positive sentiment correlation is reported to be the best indicator of a change in market conditions, and thus a regime shift. During a bull market, positive news volumes are associated with a small spread, therefore a negative correlation is present, while during a bear market the overall volume of news is positively correlated with the bonds spread. The author also reported that news sentiment is more correlated to spread volatility rather than spread difference, and this is especially true in the case of news volumes.

[Tha15] found that in the Chinese market social blog sentiment that is normally Granger caused by more technical newswire sentiment and its time-varying sensitivity has been one of the possible drivers for wild market swings in 2014/15 and that this sensitivity has decreased in June 2015 leading to a more stable stock market.

[HKGM19] studied the markets' reaction to news flow, finding a different time-varying sensitivity of prices moves to positive and negative events in different regions. The authors state that sensitivities represent a proxy for the underlying investor positioning. Some postulates identified by the Deutsche Bank team to explain muted market reactions to the news are also reported. In case of positive muted reaction the postulates are: "Holders are at their risk limits", "Existing stake reflects full conviction", "Marginal investors wary of stock being crowded, overvalued or *too consensus*" and "Early investors sell into positive news". While for negative muted reaction the postulates are: "Skittish longs already closed out positions", "Stock expensive or not available to borrow", "Existing holders believe the market has overreacted" and "Investor fatigue to the slew of poor news".

[HLGC$^+$18d] report that different factors deliver disparate performance under different macroeconomic regimes. In the paper, the result of a framework proposed by J.P. Morgan for factors timing in cross-asset risk premia was analyzed. The framework generates views for a Black-Litterman model that are translated into tactical portfolio tilts. The results showed when the risk premia were generating more value.

The non-stationarity problem evidenced by those studies is in line with the Adaptive Market Hypothesis and has, therefore, to be considered.

2.3.4. News categories and topic codes

The more advanced news analysis tools are also able to identify event types and topics discussed in the news.

[SW11] found that events aggregated by broad topics category contain little information about asset returns while considering news sentiment improves the predictive power.

As already mentioned, [BFKR13] compare the performance of stocks subdividing them into three groups. The first group is composed of stocks with no news, while the other two are composed of stocks with news. The former for stocks where the news event type can be clearly recognized, and the latter for stocks where the news is without an event type. Post-announcement (on the following days) identified news day exhibits continuation, with the largest one for the news categories: Analyst Recommendations, Deals, Employment, and Financials.

The pre-analyzed news sentiment platforms manage the news events categorization in different ways. Bloomberg and Thomson Reuters provide a list of topic codes for each article, while RavenPack provides a unique event type from a taxonomy of events types organized in a hierarchical fashion.

[Dim18] showed the existence of linear correlation among (10 minutes Pre-Opening) Bloomberg news sentiment, and also Bloomberg Twitter feeds sentiment, and contemporaneous (Open to Close) returns. In the same presentation, the author also showed that articles tagged with a wide set of multiple topic codes can be reorganized using machine-learning techniques into a more useful set of factors able to improve the impact on contemporaneous returns. As an example, the presentation shows the case of stories carrying controversial topic codes, as reported controversial and non-controversial news has a different impact on the returns. The author introduced Pi-CA, a technique to improve the selection of topic codes subsets to extract factors of interest.

Wolfe Research in [LRW$^+$19] provides an extensive study on the news event categories provided by RavenPack. In the paper, it is stated that market reaction to news and sentiment varies tremendously depending on the type of corporate event. Investors tend to overreact to bad news as litigations and layoffs, leading to a reversal post news release, and on the other side, boring news as buybacks and dividends are overlooked, leading to a momentum effect. The analysis is focused on news with high relevance for the asset (above 90 over 100), and on fresh news, with a novelty above 30 days. A relevant issue highlighted by the research is the orthogonality of the signal with respect to traditional investing factors. As highlighted, the news categories could allow better discrimination of the orthogonal signal because, for example news categorized as earnings, revenues and stock prices could already be captured by traditional factors as momentum and reversal, while underrepresented low-frequency news, as regulatory, legal, or labor issues are normally not captured by those factors. Equity analysts and credit opinionists are market movers, but these are not informative, especially if there is no change in the sentiment.

By exploiting news categories, also contradictory results can be found for the already discussed proverbial phrase “No news is good news”. The author firstly points out that news volume factors have to be adjusted for coverage, and vary by news type, and then reports that for Executive appointment and Merger & Acquisition “No news is good news” while for buybacks, revenues and earnings volume is positive.

More in detail, the author found that: for Merger & Acquisition news, positive sentiment is associated with a significant pre-announcement movement, possibly due to leakages or consecutive news, a rally on the announcement day, and underperformance of the stock post-announcement leading to mean reversal trading opportunities. Negative sentiment for Merger & Acquisition, instead, is associated with a modest previous and same-day movement and a persisting downward movement post-announcement, hinting momentum.

For buybacks news, the author found that independently of positive/negative sentiment the pre-announcement period is negative, as buybacks are normally initiated on a short-term underperformance, while, still independently of sentiment, next-day and same-day returns are positive.

For ownership news, the author found that the assets already rallied before the announcement and that the sentiment is relevant because post-announcement drift is significant and positive only for positive news.

For Rating Changes, the author found that in the case of analyst ratings there is no pre-announcement effect because there are no or very few leakages in the highly regulated sector of analysts. While there is an immediate reaction at the announcement and a modest post-announcement reaction, slightly bigger for negative news. Differently, for Rating Changes by Credit Analysts, probably due to the conservative behavior of such raters that tend to endogenously react to prices, the author found that the market reacts strongly to downgrades and negative news, overreacting to negative news, leading to a reversal after the negative event and moreover that concern on the rating can be more dangerous than an upgrade, as the investors tend to follow the “Sell first, think later” pattern.

For Earnings Guidance and Dividends, similarly to Analyst Ratings, the author found a strong pre-announcement effect and a post-announcement longer persistence for positive news.

For layoffs and legal issues, which are inherently negative, the author found that, as the companies suffer for a long time before this kind of announcement. The negative movement is distributed over the pre-announcement period and there is negative overreaction during announcement-day, leading to reversal post-announcement.

For market shares and partnership, the author found abnormal same-day returns, but limited scope for post-announcement movements.

RavenPack also provides two sector-specific categories: Bio-Pharmaceutical clinical trials and retail same-store sales. The author found that negative news for clinical trials has a strong same-day negative effect and no post-announcement effect, while for positive news there is a longer period soft-effect. For same-store sales, the author supposes that information is leaked in advance, and found a strong pre-event movement followed by a persistent positive drift post-announcement, especially for positive news.

Peter Hafez in his presentation “Reshaping Finance with alternative data” [Haf16], from the 4th RavenPack Annual Conference, shows the importance of the RavenPack taxonomy to classify different event categories, allowing for a non-uniform treatment of the sentiment. It results useful for modelling decay and for noisy event filtering. In the presentation, a different filtering technique based on temporal event chains is also introduced. Given an ongoing event, it is possible to expect the outcome of one of the possible events following in the chain, but also to filter out news related to events preceding the current event in the chain, marking those as not fresh, subsequent releases of a past event. In the presentation, it is also suggested the possibility to propagate the news effect through different kinds of dynamic networks in order to increase trading opportunities and in detail supply-chain, competitive landscape, and co-mention networks are suggested.

[HGC15] analyzed the effects of RavenPack category groups on different markets, US and European, and also for different capitalization sizes, small-caps and large/mid-caps. The results show a strong regional overlap for event groups between different sizes and a strong overlap amongst important event groups across both dimensions size and region, but also many idiosyncrasies, for instance, some event types may have momentum in a region and reversal in another. As already reported, the market response was found consistent with sentiment across a ten-day period.

The most relevant category groups in terms of market response have been identified in the US market as “Insider Trading”, “Analyst Ratings”, “Revenues”, “Earnings” and “Stock Prices” for large/mid-caps and “Acquisitions-Mergers” for small-caps. While in the European market they have been identified as “Analyst Ratings”, “Price Targets”, “Earnings” and the pair “Stock Prices” and “Revenues” for large/mid-caps and the pair “Acquisitions-Mergers” and “Dividends” for small-caps. The main differences between the two regions are that US market reacts more to “Insider Trading”, while European market to “Price Targets”, and that in the US market a short-term reversal effect is found for “Stock Prices”.

For large/mid-capitalization companies the following results were reported: for “Analyst Ratings” and “Price Targets”, a momentum effect was found, with a longer-lasting effect for the European market. For “Stock Prices”, a reversal effect was found for positive sentiment after large prices moves are reported, while negative sentiment leads to continued underperformance. For “Acquisitions-Mergers”, an 8-

10 days lagged reversal effect was found for positive news, while for negative news a continued underperformance was found only for the European market. For negative "Earnings", a linear continuation effect was found across most of the time period, while positive events are not significant. For "Revenues", negative events were found to have a stronger impact on the US market while, oppositely, on the European market, positive ones have a stronger impact. For "Insider Trading", a significant linear continuation effect was found for the US market. For "Industrial Accidents" on average, a negative impact was found, characterized by overreaction in US and underreaction in Europe leading respectively to reversal and momentum. For "Technical Analysis", a reversal signal was found, particularly after 5 days and for bullish patterns. For "Order-imbalances", prices were found to move consistently with concurrent returns, while sentiment tends to be somewhat contrarian.

While for small capitalization companies the sentiment was found to have a stronger concurrent impact and the signals are generally more persistent for small-caps, indicating less market efficiency. The US market is characterized by more negative events for "Credit ratings", "Dividends", "Equity Actions" and "Legal". For "Acquisitions-Mergers" negative sentiment was found to produce a contrary effect across all-time horizons. In the Europen market, negative "Earnings" and "Revenues" are not significant. "Industrial Accidents" were found to be not significant, or more likely were under-reported. For "Technical Analysis", a partially contrarian concurrent signal was found. Prices move with sentiment for positive events and against sentiment for negative ones.

[KKX20] introduced a new text mining methodology based on topic modelling. The methodology extracts information from news articles to predict asset returns using a three steps approach based on supervised learning. The main statistical tools applied are correlation and maximum likelihood resulting in a white box approach requiring minimal computing power. The sentiment scoring model extracted from the joint behaviour of article text and stock returns is specifically adapted to the dataset and relies on a simplified two topics model related to positive and negative returns. A measure of sentiment novelty is also extracted using cosine similarity between articles. The authors analyzed a reliable and actively monitored news source, the *Dow Jones Newswire*, and found that information is assimilated into prices with inefficient delay consistent with limits to arbitrage. The effect of news is fully reflected in prices in 4 days and is stronger and longer-lasting for fresh news, on the sell side, and for small and more volatile stocks. The resulting stock selection strategy resulted to be profitable also when transaction costs are considered.

[HGCG$^+$19] report a detailed study, by Empirical Research Partners, on big biotech and pharma industries. In these two sectors, industry-related news categories, such as patent grants, ongoing clinical trials, and FDA priority review grants, were found to have long-term effects on excess returns. Differently from others, these effects are strongly significant one year after the announcement.

[HLGC$^+$18a] report a detailed study by Citigroup's Equity Research team on CAPEX funds used by companies to manage physical assets. Highly reported CAPEX firms experience poor future stock returns. This result has been related to reports as lagging variables. The research suggests a strategy based on CAPEX announcements, rather than reported CAPEX, resulting in a positive performance with a long drift of around three months.

2.4. The uses of news

2.4.1. Univariate forecast

The news sentiment time series have a wide spectrum of applications; one of the most discussed in literature is the prediction of univariate time series regarding assets return. [Bor15] inspected correlation between jumps and news sentiment. [BCP17] use the sentiment to predict outcomes exceeding value at risk. [BCP17] found that sentiment granger causes changes in returns and volatility for commodities, in particular for gas. [BKLB14] found that browsing activity is a predictor for volumes and changes in return and volatility. [Sma16] highlighted that VIX is a better predictor than sentiment for the variance. [LXB18] estimated the elasticity in the relationship of news volumes with volatility and volumes. [IKSSS16] found that stocks emerging for a minimum period of one month of media pessimism have a good forward one-year excess returns potential.

[YMY13] and [MdBBY15] used sentiment impact to forecast assets return, volatility, volumes and liquidity. The possibility to forecast volumes and liquidity, where liquidity in trading is inspected through the bid-ask spread and the market depth, is interesting because these measures reflect and allow to monitor the conditions of the markets, and if the signal is significant can be exploited to determine if a trade can be executed in a given point of time or it is not valid. The news sentiment factor results to be not substantial in AR models for return and liquidity, while it results to be highly significant as an external factor in GARCH models for volatility prediction, leading to superior returns in trading strategies. The news sentiment impact score showed an increased predictive ability in respect to cumulated news sentiment.

[ES18b] used the macroeconomic news sentiment to model bond yield spread. The author discovered that the correlation can be used to discover early warning signs for unexpected changes in yield or structural changes visible in yield spreads. In the paper, an ARIMAX model is used to forecast the bonds spreads, the number of all news results to be particularly significant, but correlation and also residual errors change over time, as the best external variable used as predictor does. As previously reported, the author proposes to monitor correlation changes to recognize changing market conditions.

[ES17] analyzes the correlation between corporate bonds and country-specific macroeconomic news sentiment using ARIMAX models. The author found that country positive news is more effective in the economic recovery period, while country negative news predicts well during a recession. It is also found that both positive and negative improve one step ahead forecast of spread. The author inspected in detail the German parliament news sentiment, and found a strong asymmetric effect on German corporate bonds, with positive news enhancing the prediction and negative news having a limited effect. Central Bank news was also analyzed: the effect on corporate bonds is reported to be mixed, but positive news predicts well during recovery, while negative during a recession. The firm-specific news sentiment was also tested and the negative sentiment was found to be a better predictor than the positive one for corporate bonds.

In [GKG+14], [GKG+15] and [KGU+15] the authors developed an Ising model from statistical mechanic, based purely on information flow, and they applied trading strategies to the results to forecast future market index returns. They report that on the intraday scale the market is informationally efficient, due to the competition of intraday traders, while it becomes inefficient on longer timescales. They noticed that when new information is released, price change quickly, but this price-change may also be an important event that will draw media response. The original event can thus trigger a “ripple effect” of the interlinked price-change and news release, unfolding over an extended time period. In the original model, the price change is a function of the investor’s sentiment and its variation. The investor’s sentiment variation, in turn, is caused by information flow, while the information flow variation is caused by price change and exogenous news. In the most advanced model, investors with different time horizons are modeled. The four resulting strategies selected, calculated on the SPY index, outperformed the index and an active benchmark.

2.4.2. Portfolio weights estimation

The news sentiment scores have also been applied in portfolio management to enhance asset selection and weighting. In Markowitz’s theory volatility is a measure of risk. The volatility forecast based on news sentiment and volumes can therefore be used to improve the assets weighting process. In the same way, other strategies, such as the no-news vs neutral-news one, can be exploited to improve the portfolio returns.

[HS17] applied a 52-week sentiment momentum strategy with a 4-week gap based on news sentiment to improve asset allocation, highlighting the possibility to use sentiment in medium-term portfolio strategies.

In [Cre15], a Black-Litterman model is applied for portfolio optimization, in order to enhance a Markowitz framework with news sentiment, corporate social media indicators, fundamental indicators, accounting variables, and analysts’ recommendations. The author adopted news sentiment to generate views for high-frequency trading and the other indicators for medium-term asset allocation with quarterly data. The views for the Black-Litterman model are characterized by an estimation of the excess returns and a confidence level in the view imposed on the Gaussian prior distribution of the excess returns. The author found that forecast based on quarterly data of social network and fundamental indicators outperforms the market portfolio, demonstrates that news sentiment has an important high-frequency effect on returns and that in simulations news sentiment-driven portfolios outperform the market portfolio and the market index.

In [HGCD15], the authors apply thematic alphas streams (theme-based sentiment indicators) from previous research ([HGC15]) to improve equity portfolios. They combine different assets into a synthetic one for each alpha stream and then combine the alpha streams to obtain a “super-alpha” portfolio with the benefit of being able to reduce turnover through internal crossing. The portfolios have been constructed for different holding periods from one to ten days, independently selected between different alpha streams, using a different strategy to select stocks to trade. The strategy was based on ranking: long top 20% sentiment-wise stocks and short bottom 20% or just long stocks with positive sentiment and short those with negative sentiment. Different lookback windows were considered of 3 months, 6 months, and one year. The optimal lookback period was found to differ between different regions and between different sizes. US stocks require shorter windows than EU, while small-caps require longer windows than large/mid-cap ones. This effect is probably due to the amount of news needed to achieve a reasonable number of statistically significant alpha streams. The results were found to be consistent between regions and sizes. EU outperforms the US, and small-caps outperform large/mid-caps. Such results reflect the fact that the US region is considered more efficient than the EU, and market efficiency is normally higher for larger market capitalization groups. The turnover is reduced from 95% to 43-45% for small-caps and 50-53% for large/mid-caps and the strategy was found to be profitable also after applying trading costs.

[YM18] used news sentiment impact in an Enhanced Indexation framework for stock pre-filtering. The authors applied Second-order Stochastic Dominance as a criterion to enhance the returns distribution starting from a benchmark, and “volatility pumping” as money management criterion to control the maximum drawdown. The benchmark is the considered market index, the FTSE100, starting from which the framework searches for a local optimal solution. The authors compared the resulting portfolio returns, that were found to outperform the market index, with two more portfolios generated using the same technique but filtering the available stocks with different rules. The former portfolio is characterized by a rule that takes a long position only on oversold stocks, having a Relative Strength Indicator under the level of 30, and takes a short position only on overbought stocks having a Relative Strength Indicator over the level of 70. The latter portfolio, instead, is characterized by two rules: the former is the previous Relative Strength Index rule, while the latter is a rule taking long positions only on stocks with a positive news impact score and short positions only on ones with a negative impact score. The authors state that the three strategies lead to progressive improvement in the performance of a passive index fund and planned further refinements of this strategy detecting market-regimes and thus limiting the long-short partitioning.

[HLGC+18e] compared different intraday portfolio strategies exploiting *Sum Excess Sentiment Indicator* (SESI). The authors found that the performance of the strategies is strongly influenced by the large volume of news clustered in the pre-open and after-close. The results showed that the best performance is achieved with an open-to-open strategy using a 4 hours lookback window for news cumulation. Also, an 18 hours open-to-open strategy and a 10 hours close-to-close strategy are reported to represent peaks of the strategy performances. These results underline the importance of spikes in news volume at pre-open and after-close and show that these are better exploitable at market open also if some liquidity problem could arise right after the opening.

[HLGC+19] compared a portfolio strategy exploiting *Sum Excess Sentiment Indicator* (SESI) with other random and momentum-based strategies. The news sentiment strategy is based on *Extreme Sentiment Portfolio*, i.e. selecting assets with the highest sentiment strength. The news sentiment resulted to be a powerful tool for stock selection with stronger value for short-term than for long-term portfolios and higher quintiles in sentiment outperforming lower quintiles. The best results are achieved for one-day holding period. Also progressively reducing the number of selected stocks to these with the most extreme sentiment improves the returns.

2.4.3. Other applications

[CI18] applied news sentiment in order to generate indexes for geopolitical risks.

[ES18b] used the correlation between macroeconomic news sentiment and bond spreads to detect changes in market regimes and detect regime shifts.

[KJF17] applied the social anomaly score (SAS) to forecast volatility. The SAS indicator was developed by PsychSignal and analyzes messages from Twitter and StockTwits, reflecting each stock symbol activity level. The authors calculated the difference between a capitalization-weighted SAS of the S&P500 stocks and the SAS of SPY, an ETF on the S&P500 index. The difference was used to implement a trading strategy in order to decide to take a long or short position on exchanges-traded products available from the CBOE market volatility index (VIX). The strategy was reported to outperform the considered benchmark also after fees and trading costs.

[IKSSS16] used SEC/EDGAR fillings to estimate asset returns similarity.

2.5. Our contribution

To the best of our knowledge, our contribution to the existing literature in this field regards mainly the two models presented in chapters 4 and 5.

In chapter 5 we analyzed the effects of the cross-sectional news factors over three different baseline models. The cross-sectional news factors analyzed are extracted from the news category groups present in the RavenPack news taxonomy and over different cumulation periods. The effects of the cross-sectional news factors are then tested on baseline models exploiting three different criteria: the Sharpe Ratio, the Second-order Stochastic Dominance, and the Scaled Second-order Stochastic Dominance.

In chapter 4 instead, we verify the performances of the pre-opening news sentiment and volume associated with each one of the news category groups separately. The effects of the news are evaluated in a naive beauty context model. The effects are analyzed taking long or short positions on the market according to the news. The effect of positive only and negative only sentiment is also considered in the analysis of the strategies.

3. Considered Data

The aim of this research is to inspect the possibilities to enhance asset allocation and portfolio management using information extrapolated from relevant firm's specific public news. To achieve this result it is of interest to understand which kind of news affects market prices on an intraday, but also on a multiple-day time scale and how every single piece of news has to be treated to obtain meaningful indicators for the desired task.

Firm's specific news pieces are supposed to affect stock returns because news reports events that could affect stock fundamentals or influence investor's behaviour. At the same time markets could not immediately discount all information correctly.

As previously reported, the information is widespread by news whose effect can be measured by different indicators. Some indicators are more easily accessible than others, and not all are always available. Some examples of these indicators are the sentiment, volume and novelty of the news, but also the level of attention generated by the news that can be proxied by the number of web pages views or the volume of search engines queries.

The data analyzed in this work were kindly provided by OptiRisk Systems LTD.

3.1. The data sources

3.1.1. Considered markets and periods

The provided database contains information about companies belonging to different market indexes. The lists of the market indexes components cover different geopolitical areas. The database reports data about the S&P 500 index components and US MSCI BARRA factors for the US area. MSCI factors are generated considering returns of subsets of stocks characterized by common qualities, such as momentum, low volatility, low leverage, value, quality, or earnings. For the Euro area, the database reports the components of EURO STOXX 50, DAX 30, and FTSE 100 (still in EU in the considered period). It also reports the components of NIFTY 50 for India, Hang Seng 50 for Hong Kong, and Nikkei 225 for Japan.

For each one of the indexes, the corresponding implied volatility index is also reported. The implied volatility index, also known as fear index, reports an aggregated measure of the volatility implied by the index related call and put options expiring in the near future. The reported value is supposed to describe the near future volatility (standard deviation), also if it is bounded to the forecasting abilities of the techniques used to estimate option prices and could also be affected by the liquidity of the underlying stocks. These are named VIXFUT, VIXFTSE, VSTOXX, VNIFTY, VHSENG, VNIKKEI, and VDAX.

In this work, the research interest is focused mainly on the EURO STOXX 50 market index components, while some studies are conducted on the DAX 30 and the S&P 500 components also for comparison and confirmation of the results on the EURO STOXX 50. The EURO STOXX 50 is considered of particular interest because it is less studied than others, especially compared to S&P 500.

The database reports information regarding market data and news sentiment covering the period from January 2005 to May 2018. The considered period is extended enough to allow for an evaluation of the strategies in different situations and under different market regimes. Also when a pre-calculation period of 3 years is consumed by a 3 years rolling window, used to consider the market changing dynamics, 10 years of data lasts for results evaluation.

3.1.2. Market data

The most used index, with its components, in this work is the EURO STOXX 50. The 68 stocks belonging to the EURO STOXX 50 market index at least once in the period between 2005 and 2018 are considered. The components of this index are used because European markets are reputed less efficient than US stock market even if they are not strongly affected by news sentiments in particular periods as, for example, the Chinese stock market (see [Tha15]). The log-returns on daily adjusted closing prices are considered. The considered period spans over more than 13 years, from January 2005 to May 2018.

The database reports for each stock many informations. An ID representing each stock, the market index to which it belongs, and the country in which it is listed are reported. For each stock, the opening, closing, high, and low (OCHL) adjusted prices and the traded volume are reported for each trading day of the considered period. For each one of the indexes in addition to OCHL prices and volumes also the OCHL prices of a futures contract on the index, and of the implied volatility index are reported.

Pieces of information regarding the listing periods of stocks inside indexes are reported in the database with a daily precision for each stock and index. Anyway, these data are not directly taken into account in this work. The stocks are considered also when out of the market index if the data is available in the provided database. For the researches regarding daily estimates, no other market information about stocks is considered from the database except for adjusted closing prices, while for intraday estimates of pre-opening effects adjusted opening and closing prices are considered.

To achieve a higher level of accuracy and to overcome the presence of some missing data the stock's adjusted opening and closing price time series have been updated using these provided by the Refinitiv Eikon service available from the university library.

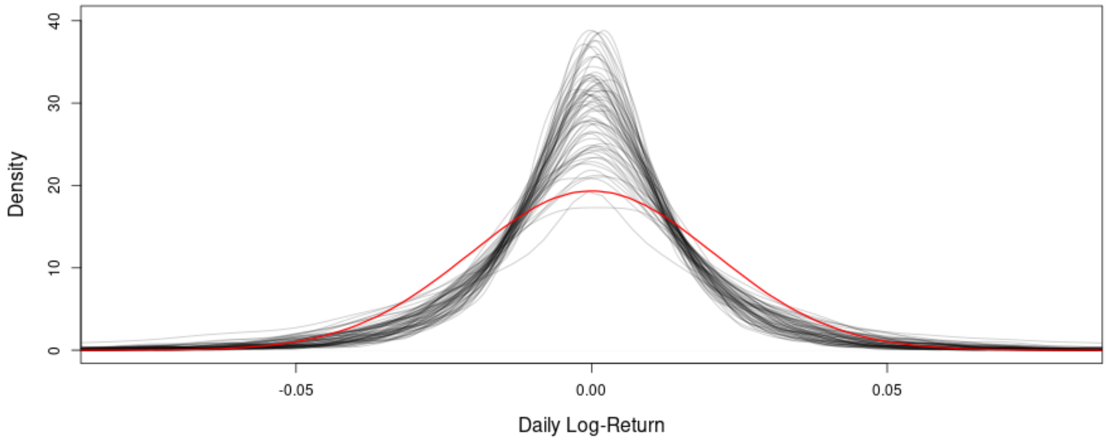

Figure 3.1.: The figure shows the density of the daily log-returns distributions of the 68 stocks belonging to the EURO STOXX 50 market index in black and a Normal distribution for comparison in red. The mean and standard deviation of the Normal distribution in red are set equal to the mean of the means and the mean of the standard deviations of the stocks.

In figure 3.1 we show the plots of the daily log-returns distributions of the considered stocks and the relative average gaussian distribution in red. The stock distributions appear quite similar between them with few exceptions, where the main difference is how much each distribution is peaked in 0, and thus probably related to its total variance and fat tails. Comparing the stock distributions with the gaussian reference we can see that in general, these are more peaked in 0 than the reference. These also show inferior variance compared to the reference in the neighbourhood of the inflection point on the slope of the distribution curve but then are characterized by fatter tails than the reference.

In table 3.1 we show the statistics about stocks. The majority of stocks present a similar number of samples, due to the presence of price in the database for these stocks over the whole considered period during trading days. Only 3 companies on 68 have a sample size under 90% of the maximum sample size. The average mean is positive, the average standard deviation is 0.021, the average skewness is negative, and almost all the distributions are leptokurtic. Fortis N.V. and AIB Group Plc show the highest standard deviation, the most negative skewness, and the highest kurtosis. ASML Holding N.V. obtains the highest average return in the period.

The formulas used in the tables to calculate the skewness, third standardized moment, and the zero centered kurtosis, zero centered fourth standardized moment, of the distributions are respectively:

$$\tilde{\mu}_3 = \frac{E[(x - E[x])^3]}{E[(x - E[x])^2]^{3/2}} \qquad \tilde{\mu}_4 = \frac{E[(x - E[x])^4]}{E[(x - E[x])^2]^2} - 3$$

Table 3.1.: This table reports statistics regarding the companies belonging to the EURO STOXX 50 market index at least once in the considered period between 2005 and 2018. The table reports the name of the company, shortened when needed, the numerosity of the sample and the main statistical parameters.

Company	#Samp.	Mean	Std.Dev.	Skewness	Kurtosis
ABN AMRO GRO...	661	0.000297	0.0171	-1.1639	9.31
AEGON N.V.	3499	-0.000162	0.0279	0.1533	15.49
AIB GROUP PLC	3499	-0.001976	0.0525	-1.5525	30.83
AIR LIQUIDE ...	3499	0.000292	0.0143	-0.0302	4.23
AIRBUS SE	3499	0.000439	0.0219	-0.8565	13.49
ALCATEL-LUCENT	3499	-0.000379	0.0301	-0.3016	6.99
ALLIANZ SE	3499	0.000174	0.0198	0.5417	14.73
ALSTOM S.A.	3499	0.000360	0.0232	0.0962	5.23
ANHEUSER-BUS...	3499	0.000432	0.0182	-0.8781	16.34
ARCELORMITTA...	3499	-0.000233	0.0298	-0.1537	4.66
ASML HOLDING...	3499	0.000755	0.0193	0.2447	2.90
ASSICURAZION...	3499	-0.000121	0.0175	-0.2204	6.62
AXA S.A.	3499	0.000055	0.0248	0.2305	10.05
BANCO BILBAO...	3499	-0.000157	0.0210	0.1246	7.19
BANCO SANTAN...	3499	-0.000048	0.0217	-0.1219	9.92
BASF S.E.	3499	0.000338	0.0176	0.0039	8.67
BAYER AG	3499	0.000430	0.0172	-0.1053	4.06
BAYERISCHE M...	3499	0.000272	0.0195	0.0754	4.69
BNP PARIBAS ...	3499	0.000007	0.0246	0.1905	9.47
CARREFOUR S.A.	3499	-0.000202	0.0181	-0.1900	4.13
COMPAGNIE DE...	459	0.001091	0.0224	0.2938	0.91
CREDIT AGRIC...	3499	-0.000159	0.0266	0.2557	6.77
CRH PLC	3499	0.000162	0.0224	-0.2115	4.45
DAIMLER AG	3499	0.000161	0.0209	0.2277	8.03
DANONE S.A.	3499	0.000203	0.0142	-0.0601	4.09
DEUTSCHE BAN...	3499	-0.000500	0.0253	0.2684	8.89
DEUTSCHE BOE...	3499	0.000480	0.0209	0.0168	6.33
DEUTSCHE POS...	3499	0.000185	0.0177	-0.3613	11.35

DEUTSCHE TEL...	3499	-0.000069	0.0155	-0.0674	9.33
E.ON SE	3499	-0.000227	0.0185	-0.1386	6.58
ENDESA S.A.	3499	0.000285	0.0157	-0.1296	5.53
ENEL S.P.A.	3499	-0.000081	0.0171	-0.2887	6.96
ENGIE S.A.	3366	-0.000226	0.0182	0.3309	11.70
ENI S.P.A.	3499	-0.000043	0.0170	0.3013	9.78
ESSILOR INTE...	3499	0.000410	0.0136	0.4008	6.63
FORTIS N.V.	3499	-0.000405	0.0377	-18.3252	726.98
FRESENIUS SE...	3499	0.000612	0.0162	-0.0352	3.58
IBERDROLA S.A.	3499	0.000206	0.0174	0.2750	12.40
INDUSTRIA DE...	3499	0.000531	0.0169	0.1804	4.10
ING GROEP N.V.	3499	-0.000095	0.0294	-0.0685	16.27
INTESA SANPA...	3499	-0.000078	0.0259	-0.4611	8.87
KONINKLIJKE ...	3499	0.000343	0.0146	-0.1712	5.31
KONINKLIJKE ...	3499	0.000223	0.0178	-0.0399	4.47
L'OREAL S.A.	3499	0.000377	0.0144	0.2342	5.70
LAFARGE S.A.	3499	-0.000019	0.0221	0.0175	4.64
LVMH MOET HE...	3499	0.000513	0.0174	0.1350	5.45
MUENCHENER R...	3499	0.000192	0.0153	0.1046	8.38
NOKIA CORP.	3499	-0.000255	0.0254	-0.3157	12.17
ORANGE S.A.	3499	-0.000148	0.0157	0.0836	3.98
RENAULT S.A.	3499	0.000081	0.0251	-0.2137	4.83
REPSOL S.A.	3499	0.000058	0.0190	-0.2623	5.85
RWE AG	3499	-0.000219	0.0197	-0.1305	6.38
SAFRAN S.A.	3499	0.000544	0.0206	-0.2233	4.63
SANOFI S.A.	3499	0.000030	0.0154	-0.1601	6.10
SAP SE	3499	0.000315	0.0148	-0.4767	12.69
SCHNEIDER EL...	3499	0.000308	0.0205	0.0260	5.30
SIEMENS AG	3499	0.000177	0.0183	-0.1113	15.40
SOCIETE GENE...	3499	-0.000173	0.0292	-0.1393	7.05
SUEZ	2574	-0.000170	0.0185	-0.2897	7.47
TELECOM ITAL...	3499	-0.000439	0.0214	-0.0601	4.25
TELEFONICA S.A.	3499	-0.000130	0.0154	-0.4499	9.11
TOTAL S.A.	3499	0.000079	0.0159	0.1200	6.73
UNIBAIL-RODAMCO	3499	0.000189	0.0168	-0.3974	6.30
UNICREDIT S....	3499	-0.000600	0.0297	-0.1884	6.86
UNILEVER N.V.	3499	0.000324	0.0171	0.2297	2.55
VINCI S.A.	3499	0.000357	0.0184	0.2669	8.21
VIVENDI	3499	-0.000024	0.0159	-0.0380	4.75
VOLKSWAGEN AG	3499	0.000552	0.0241	-0.8235	11.81

3.1.3. News data

The analyzed news data stream source is a pre-filtered version of the RavenPack News Live Stream. The RavenPack News Live Stream is a news stream that includes news published by reputable content sources in English. Publishers include Dow Jones Newswires, the Wall Street Journal, Direct Regulatory and PR feeds and over 19000 other traditional and social media sites.

The news data stream from RavenPack is composed of a separate SQLite Database Table for each considered company, where the company name is encoded using proprietary identification numbers.

SQL 1

```
select date, hour, categoryGroup, relevance, ens, ess from news0820E6
```

	date	hour	categoryGroup	relevance	ens	ess
480	20110817	234340	technical-analysis	100	1.02485	0.45
481	20110823	234841	technical-analysis	100	16.01095	-0.43
482	20110825	11005	analyst-ratings	100	365.0	0.94
483	20110825	11048	price-targets	100	365.0	0.78
484	20110830	70425	equity-actions	100	365.0	0.0
485	20110830	72615	equity-actions	100	0.01516	0.0
486	20110831	1716	stock-prices	100	92.92102	-0.3

Figure 3.2.: Example of SQLite Database Table for a company (with ID:0820E6). The table reports the entries from RavenPack News Live Stream used in this study. For each piece of news different fields are reported. Date and Hour allow to exactly locate the news in time. The category group represent a macro categorization of the news in the RavenPack taxonomy. Relevance, event novelty score (ENS), and event sentiment score (ESS) are fields extracted from the news article using natural language processing techniques.

The 68 news tables for EURO STOXX 50, characterized by the company ID, share the same repetitive structure composed of multiple fields. The dataset includes the exact date and time for each news, allowing to correctly locate the news in time and discount its effects through the indicators.

The news category and category groups are two fields in the tables that position each news event inside the RavenPack events taxonomy, a hierarchical representation of similar news events. The category group represents a macro categorization in respect to categories, where many argument related categories are grouped together. The RavenPack events taxonomy contains more than 2000 categories and more than 50 category groups. In the whole database news pieces related to 48 category groups and 1072 categories are reported. Following the work of [HGC15], in the analysis to generate news streams of interest the category group field has been used to group at an appropriate level the different typologies of events. In the considered period, 38 different category groups are present in the database for EURO STOXX 50.

The news relevance is a field that for each news event specifies the relevance of that particular piece of news for the stock considered by a particular table, the value is in the range between 0 and 100, with 100 meaning that the stock is strongly related to the news, playing a central role in the story, or is mentioned in the headline. The pre-filtered database contains only news with a relevance score of 100.

The news event novelty score (ENS) is a field specifying the number of days in the past without similar news. This field takes values in the range between 0 and 365 and has been used in the analysis to weight fresh news against stale news. See for example 3.2. In the fifth column of the table, we can see the number of days without similar news for each news referring to the considered company.

The news event sentiment score (ESS) is a field in the tables specifying the short-term financial or economic impact of the news. The score is determined by systematically matching stories typically categorized by RavenPack financial experts. The score is derived using machine learning techniques from a training set where financial experts classified entity-specific events as positive or negative. The original value range between 0 and 100, while in this database pre-rescaled values are already present with values ranging between -1 and 1. See for example figure 3.2. In the sixth column of the table, we can see the sentiment score relative to each news referring to the considered company.

In the database two more news related sentiment scores are provided but are not exploited in this work:

- The Composite Sentiment Score (CSS) is a sentiment score relative to the terms used in the news article, determined by examining emotionally charged words and phrases and by matching stories typically rated by experts as having short-term positive or negative share price effect.
- The News Impact Projections (NIP) score is an intraday impact score calculated considering the degree of impact over the following 2 hours on large-capitalization stocks.

In the considered period 817277 stock-related news events are present for EURO STOXX 50. Each stock has between 0 and 57704 pieces of news, on average there are 12019 news pieces for each stock. Each category group has between 1 and 130674 news, on average there are 21507 news pieces for each category group.

In table 3.2 we show the statistics about category groups. 19 category groups on 38 have less than 3536 news pieces, which mean less than one news piece per quarter for each stock, and therefore these are probably not suitable for cumulated sentiment indicators over long periods. 3 category groups have only news pieces with neutral sentiment (the average sentiment is exactly 0), these category groups are not suitable for cumulated news sentiment indicators, but only for news volumes indicators. Almost all the category groups relevant for cumulated indicators over long periods have an average novelty greater than 60 days, an higher novelty score gives chances to the news pieces to have more impact on stock prices.

Table 3.2.: This table reports for each category group present in the database, the number of categories belonging to it, the number of news pieces present for the EURO STOXX 50 in the considered period, the average event novelty score (ENS) and the average event sentiment score (ESS) of these news pieces.

Category Group	#Category	#News	E[Novelty]	E[Sentiment]
acquisitions-mergers	106	96726	68.78	0.4638
analyst-ratings	16	28545	202.25	0.1225
assets	35	18955	83.90	0.0561
balance-of-payments	16	158	81.36	0.0204
bankruptcy	7	119	31.43	-0.1958
civil-unrest	6	98	59.11	-0.0706
corporate-responsibility	2	653	77.90	0.1755
credit	25	6670	85.02	0.1201
credit-ratings	47	16685	152.20	-0.0486
crime	11	140	70.54	-0.0733
dividends	15	9071	206.75	0.0852
domestic-product	14	36	25.32	0.0000
earnings	161	130674	218.57	0.1314
equity-actions	134	43151	102.50	0.1707
exploration	4	961	25.59	0.0924
government	1	13	12.44	-0.0353
health	2	61	4.03	-0.0329
housing	5	1	5.37	0.0000
indexes	7	97	92.69	0.0727
industrial-accidents	25	916	54.64	-0.1433
insider-trading	7	526	166.26	0.1252
investor-relations	5	6053	102.11	0.0000
labor-issues	48	52651	73.54	0.0858
legal	45	24983	96.80	-0.3170
marketing	8	25172	157.01	0.1157
order-imbalances	8	2244	74.72	0.0327
partnerships	6	29139	69.36	0.4552
price-targets	7	15552	194.97	0.0666
products-services	139	129205	71.74	0.3239
regulatory	18	4305	100.99	-0.3115
revenues	56	61341	204.33	0.1722
security	13	287	61.28	-0.2666
stock-picks	8	434	184.01	0.3902

stock-prices	6	65070	45.60	0.0120
taxes	3	12	12.13	0.0371
technical-analysis	10	46277	21.93	0.0304
transportation	3	112	0.59	-0.0064
war-conflict	26	184	66.46	-0.0577

The news-event sentiment score is a granular score and it does not seem to be normally distributed when considered separately for each category group. Some category groups for example, as reported by [LRW+19], are characterized only by negative news.

3.1.3.1. Extraction techniques for sentiment indicators

Sentiment indicators are extracted by RavenPack from reputable content sources. The sentiment indicators extraction is an automatized process. The extraction process relies on advanced natural languages processing techniques and a cloud server infrastructure.

There are two main sentiment indicators extracted, Composite Sentiment Score (CSS) and Event Sentiment Score (ESS). The latter, ESS, is favoured over the former, CSS, by RavenPack itself. A comparison between the two shows that, also if the two have a good level of correlation (32.6% correlation and 43.7% Spearman's rank correlation), many news classified by CSS as neutral can be further classified as positive or negative using ESS.

The CSS is a score derived from the other 5 sentient scores, avoiding conflicts between them. These five scores are: PEQ, BEE, BMQ, BAM, BCA. PEQ is a classifier for Global Equities news, and BEE is a classifier for Earnings Evaluation. Both classifiers are dictionary-based, and assign a sentiment score depending on particular emotionally charged words or phrases in the news article. The sentiment scores are based on RavenPack's Traditional Methodology. BMQ is a classifier for commentary and editorials on global equity markets, BAM is a classifier for news regarding acquisitions, mergers, and takeovers, while BCA is a classifier specialized in reports on corporate action announcements. The three classifiers rely on scores assigned to the news by financial experts. The sentiment scores are based on the RavenPack's Consensus Methodology, and for BAM and BCA, have been trained on stories that lead up to a pre-identified related event.

The ESS is a granular score measured by various proxies from news matching stories categorized by financial experts having a short-term positive or negative share price impact. The score regards the main event discussed in a story and is assigned only for events matched by one category in the RavenPack Events Taxonomy. The score trained on the price impact is given by the category of the event plus a limited effect of financial figures, and analyst ratings directional language and represents a stylized sentiment view on a particular event.

The RavenPack Event Taxonomy is articulated onto 4 levels: 5 topics, 51 groups, 412 types, and 2064 categories. The classifier is able to detect companies also mentioned using short name, long name, abbreviations, security identifier, subsidiary information, and up to date corporate action data. At the category level in the taxonomy, differently from the type level, also the role of the company in the event could be identified if present (i.e. the acquirer for an acquisition is categorized under category “acquisition-aquirer”).

RavenPack also associates two more scores to news events for a company: The news Relevance and the news Event Novelty Score (ENS). When a company links to a category given its role, the news receives for the company a relevance score of 100. News events significantly relevant for a company receives a relevance score of 75 or above. News events receive a relevance score of 0 for a company when it is only passively mentioned in the news article. The ENS represent the number of days elapsed without similar news, to achieve this result RavenPack tries to group together news related to the same event.

The News Impact Projection score (NIP) instead is a sentiment score indicating the expected impact of the news on a stock price volatility in the two hours following the announcement. The classifier is trained on intraday market data on the 100 most capitalized stocks of the S&P500.

Three more indicators extracted by RavenPack but not present in the database are: BER, ANL-CHG and MCQ. BER is a classifier specialized in news stories about earnings releases and is based on RavenPack’s Consensus Methodology. ANL-CHG is a classifier assigning a granular score to news regarding recommendations by an analyst in terms of stock upgrades and downgrades. MCQ is a multi classifier for equities. As for CSS, the score is derived from other scores, in detail BMQ, BEE, BCA, and ANL-CHG. The score is assigned when a company is mentioned in the headline and tagged with sentiment from one of the four classifiers. The score is based on the tone towards the most relevant company mentioned in a story.

RavenPack also classifies news into different news types having a different market impact: Hot News Flash, News Flash, Full Article, Press Release, and Tabular Material.

The news article analysis performed by RavenPack exploit advanced Natural Language Processing (NLP) techniques (see [MS99] and [Gol16]). The analysis only takes into account reputable content sources in English, this choice probably reflects the necessity to limit ambiguity, reduce irony, avoid jokes and limit the presence of noisy unrelated articles.

The RavenPack Taxonomy is based on an ontology of events, that allows the matching of the events category and roles on the news articles. An advanced Named Entity Recognition (NER) is present, capable of recognize companies, peoples, products, places, organizations, currencies and commodities (see [RR09] and [ML03]). The NER is also able to identify a company using related information. To allow the categorization of the news articles also matching the role of the companies in an event an advanced Topic Detection algorithm should be used (see [BGHM17] and [Ble12]), and probably Shallow Parsing techniques applied to match the roles.

3.2. Data pre-processing

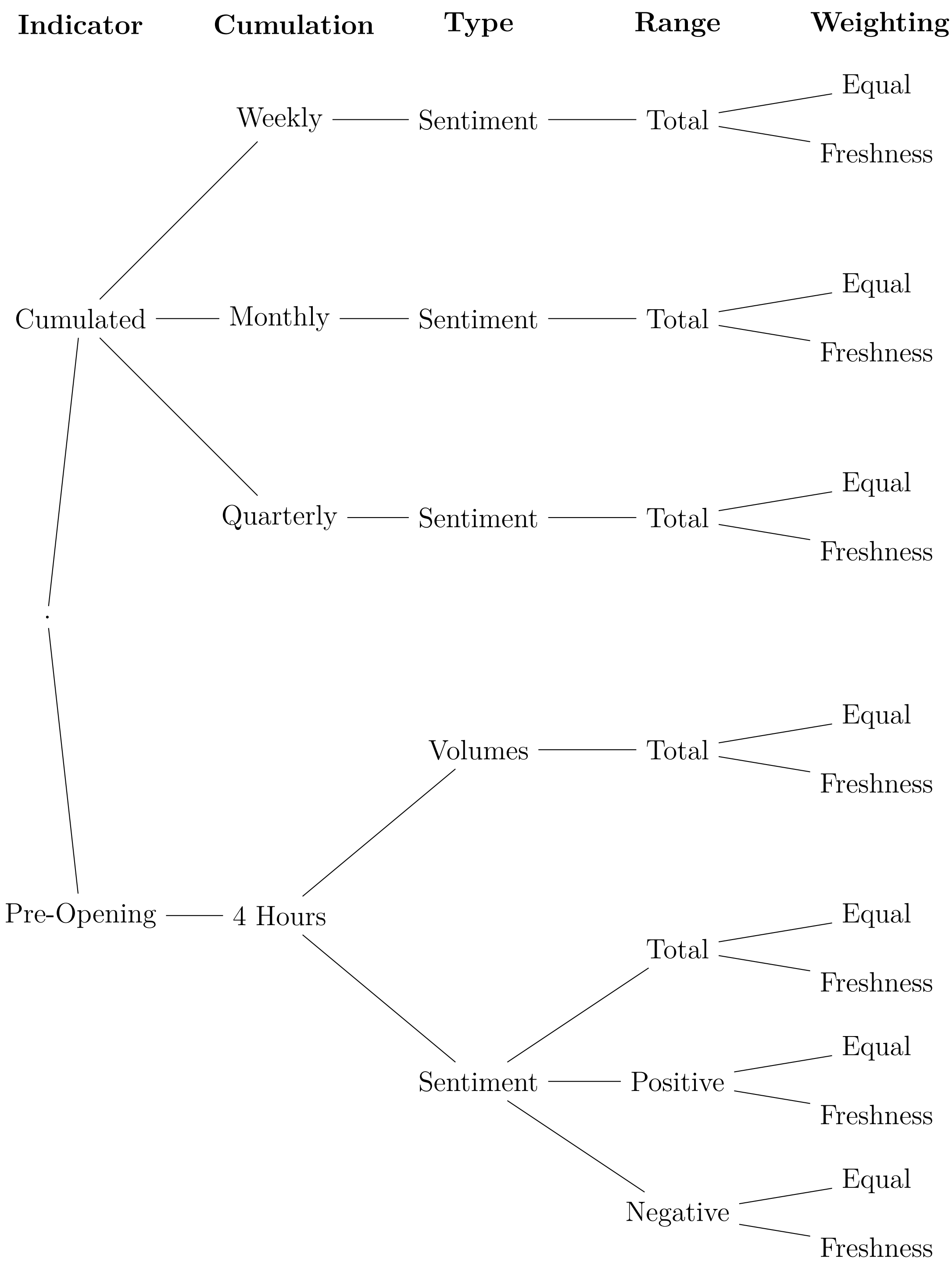

Figure 3.3.: The figure shows a tree representing the generated indicators for different kinds of cumulation periods and types of indicators.

In this section, the sentiment indicator construction process is described. In figure 3.3 a tree of the different indicators used in this research is presented. The first branch differentiates the type of indicator, between cumulated and pre-opening. The successive branches differentiate the type of cumulation, between quarterly, monthly, weekly and 4-hours, the type and range and the weighting technique applied.

3.2.1. Relevance filtering

News with relevance less than 100 was filtered out, keeping only news mentioning the relative stock in the title. The filtering was mandatory because the database from OptiRisk Systems LTD was already pre-filtered, but it is also in line with what was done by [HGC15]. Other authors filter the database using a threshold of 75, including also highly relevant news not mentioning the stock in the title. The applied filtering ensure the relevance of news for the stock, that appearing in the title, or playing a central role, could be more easily discovered, and guarantees a higher level of attention. For example, institutional investors could rely for their choices on news feeds where only the news title appears in a long list, and then they could be more affected by this kind of news.

3.2.2. Freshness indicator

An indicator of each piece of news freshness was calculated using a function of its novelty (number of days without similar news), gradually penalizing news more recent than 90 days. The penalization is calculated by applying a sigmoid function to the rescaled novelty of the news. This approach is in line with the analysis of [LRW$^+$19] and of [HLGC$^+$18b] where only news with a novelty respectively greater than 30 and 90 days was considered:

$$\upsilon_i = \frac{1}{1 + e^{\frac{200-3\eta_i}{30}}}$$

where:
η_i is the number of days without similar news (ENS) for the news i and
υ_i is the freshness weight for the news i

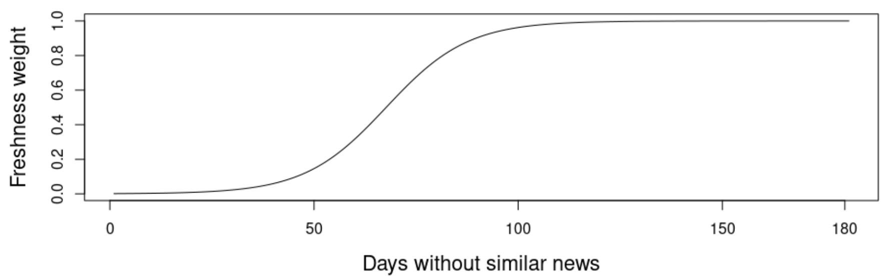

Figure 3.4.: Plot of the freshness weight corresponding to a given number of days without similar news (novelty) calculated applying the previously introduced function.

The plot, presented in figure 3.4, in line with other researches, shows a clear transition between 30 and 90 days from underweight to fully weighted.

Is of interest to analyse the freshness of each news piece because repetitive stale news could have a different lower impact on the market prices, while first time seen news could lead to stronger reactions.

3.2.3. Pre-opening indicators

News related indicators are complex to handle because data appear to be highly sparse and on a continuous time frame. To overcome the sparsity problem, the data are cumulated on a 4-hours pre-opening time period. The data cumulation process is different from a data averaging process. The cumulation has been found to better model news effects because it jointly models the news sentiment and the news volume effects. Different indicators reporting sentiment are generated for each category group and different kinds of cumulation. The considered kinds of cumulation are mainly total sentiment, total fresh sentiment, total volume and total fresh volume. Negative sentiment, negative fresh sentiment, positive sentiment and positive fresh sentiment have also been generated, but only to better inspect the effects of news sentiment.

The news cumulation process τ relative to each stock j and for each category group g starts every day at 5 am and ends every day at 9 am before the market opens.

The pre-opening cumulated indicators are defined as:

$$\forall\ j \in \mathbb{S}, \forall\ c \in \mathbb{C}$$

$$\forall\ d\ \in\ \{2005/01/01, \ldots, 2018/05/31\}$$

$$\mathbb{A} = \{a | C_a = c \wedge D_a = d \wedge 5am < H_a < 9am\}$$

$$\tau = \sum_{i \in \mathbb{A}} \kappa_i$$

$$\tau_f = \sum_{i \in \mathbb{A}} \kappa_i \upsilon_i$$

$$\psi = \sum_{i \in \mathbb{A}} 1$$

$$\psi_f = \sum_{i \in \mathbb{A}} \upsilon_i$$

$$\tau^+ = \sum_{i \in \mathbb{A}} \max(0, \kappa_i)$$

$$\tau_f^+ = \sum_{i \in \mathbb{A}} \max(0, \kappa_i \upsilon_i)$$

$$\tau^- = \sum_{i \in \mathbb{A}} \min(0, \kappa_i)$$

$$\tau_f^- = \sum_{i \in \mathbb{A}} \min(0, \kappa_i \upsilon_i)$$

where:
$\mathbb{S}$ is the set of available stocks,
$\mathbb{C}$ is the set of category groups,
$\mathbb{A}$ is the set of news pieces to consider in the cumulation,
C_a is the category groups of news piece a,
D_a is the date of news piece a,
H_a is the hour of news piece a,
κ_i is the sentiment (ESS) of news piece i,
τ is the cumulated news sentiment for stock j and category group c at day d,
τ_f is the cumulated fresh news sentiment for stock j and category group c at day d,
ψ is the cumulated news volume for stock j and category group c at day d,
ψ_f is the cumulated fresh news volume for stock j and category group c at day d,
τ^+ is the cumulated positive news sentiment for j and c at day d,
τ^- is the cumulated negative news sentiment for j and c at day d,
τ_f^+ is the cumulated positive fresh news sentiment for j and c at day d,
τ_f^- is the cumulated negative fresh news sentiment for j and c at day d.

After the data pre-processing step, a multitude of indicators time-series is generated for each stock. More than 20’000 news related time-series are generated just for the EURO STOXX 50.

In figure 3.5 we show the distribution having an extremely strong peak in 0 given by days without news and long fat tails, given by days with higher news volumes.

In table B.1 we show the statistics for 4-hour pre-opening cumulated news sentiment. Some category group is characterized by the prevalence of positive sentiment. “Acquisition-mergers” and “earnings” show a positive deviation. The skewness varies

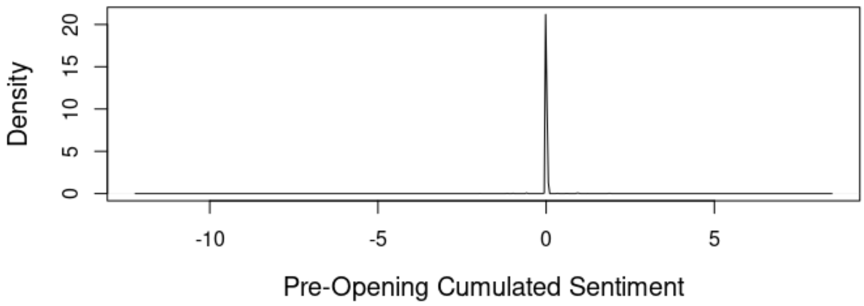

Figure 3.5.: Distribution of 4 hours pre-opening cumulated news sentiment indicator for “Analyst Rating” category group.

between category groups underlying asymmetries in the cumulated sentiment. All the category groups sentiment distributions are extremely leptokurtic with extremely high kurtosis. The high kurtosis is also emphasized by the strong peak in zero, given by days without news. There are no skewness and kurtosis for “domestic-product” and “invesor-relations” that are characterized only by neutral sentiment, for “government” and “order-imbalances” that have no-news releases in the considered 4 hours pre-opening time-window, and for “housing” that has just 1 news pieces.

In table B.2 we show the statistics for 4 hour pre-opening cumulated fresh news sentiment. The reduced means and standard deviations, compared to the previous table, show the effect of freshness weighting of news pieces. The cause of this effect is the reduced volumes of news given by the weighting process.

In table B.3 we show the statistics for 4 hour pre-opening cumulated news volumes. The means show the average news volumes for each category group. The standard deviation reflects the variability in pre-opening news volumes. The positive skewness and high kurtosis describe the shape of the distributions and reflect the characteristics of pre-opening news volumes peaks. There are no skewness and kurtosis for “government” and “order-imbalances” that have no-news releases in the considered 4 hours pre-opening time window, and for “housing” that has just 1 news piece.

In table B.4 we show the statistics for 4-hour pre-opening cumulated fresh news volumes. The reduced means and standard deviations, compared to the previous table, show the effect of freshness weighting of news pieces.

Cumulated news sentiment indicators are known to be non-normally distributed. These indicators are often characterized by a gaussian-like shape centred near 0 but also by fat-tails and a very strong peak in 0, given by days without news.

3.2.4. Cumulated indicators

News related indicators are complex to handle because data appear to be highly sparse and on a continuous time frame. To overcome the sparsity problem, the data is cumulated on weekly, monthly and quarterly time frames and different indicators are generated for each category group reporting sentiment. The data cumulation is also suitable for the processing because the news effect could not be discounted by the market immediately and at once, but instead, it could take place over a period of time in which uncertainty is removed. The data cumulation process is different from a data averaging process. The cumulation has been found to better model news effects because it jointly models the news sentiment and the news volume effects. The news cumulation process τ relative to each stock j and for each category group g (following [HGC15]) starts and ends every day half an hour before the market closes. At that moment, as reported by other studies, enough liquidity should be present in the market in order to allow the portfolio rebalancing, and the price should reasonably reflect the closing price.

The cumulated indicators are defined as:

$$\forall\, j \in \mathbb{S}, \forall\, c \in \mathbb{C}, \forall\, p \;\in\; \mathbb{P} = \{7, 30, 90\}$$

$$\forall\, d \;\in\; \{2005/01/01, \ldots, 2018/05/31\}$$

$$\mathbb{A} = \{a | C_a = c \wedge D_a \in \{d, \ldots, d - p + 1\}\}$$

$$\tau = \sum_{i\, \in\, \mathbb{A}} \kappa_i$$

$$\tau_f = \sum_{i\, \in\, \mathbb{A}} \kappa_i \upsilon_i$$

where:
$\mathbb{S}$ is the set of available stocks,
$\mathbb{C}$ is the set of category groups,
$\mathbb{P}$ is the set of news cumulation period lengths,
$\mathbb{A}$ is the set of news pieces to consider in the cumulation,
C_a is the category groups of news piece a,
D_a is the date of news piece a,
κ_i is the sentiment (ESS) of news piece i,
τ is the cumulated news sentiment for stock j and category group c at day d with cumulation period p,
τ_f is the cumulated fresh news sentiment for stock j and category group c at day d with cumulation period p.

After the data pre-processing step, a multitude of indicators time-series is generated for each stock for weekly, monthly and quarterly cumulated news. More than 7500 news related time-series are generated just for the EURO STOXX 50.

Cumulated fresh news sentiment indicators are known to be non-normally distributed. These indicators are often characterized by a gaussian-like shape centred near 0 but also by fat-tails and a very strong peak at 0, given by days without news.

In figure 3.6 we show the distribution having a gaussian-like shape, a strong peak in 0, due to periods or stocks without news for the category group, and long fat tails. Some secondary small noise-like peaks are present, probably due to sentiment granularity related effects.

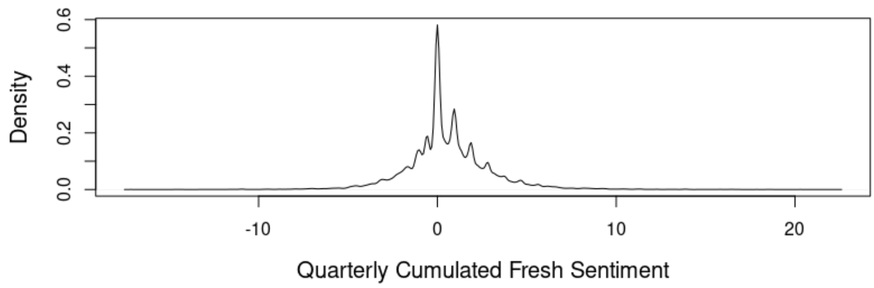

Figure 3.6.: Distribution of quarterly cumulated fresh news sentiment indicator for "Analyst Rating" category group.

In table B.5 we show the statistics for monthly cumulated news sentiment. Some category group is characterized by the prevalence of positive or negative sentiment. "Acquisition-mergers", "earnings", "partnerships", and "revenues" show a positive deviation, while "legal" and "regulatory" show a negative one. The standard deviation in cumulated indicators remains low (under 0.2) for category groups with an insufficient number of news pieces. The skewness varies between category groups underlying asymmetries in the cumulated sentiment. All the category groups sentiment distributions are leptokurtic with high kurtosis. The high kurtosis is also emphasized by the strong peak in zero, given by days without news. There are no skewness and kurtosis for "domestic-product" and "invesor-relations" that are characterized only by neutral sentiment, and for "housing" that has just 1 news piece.

In table B.6 we show the statistics for monthly cumulated fresh news sentiment. The reduced means and standard deviations, compared to the previous table, show the effect of freshness weighting of news pieces. The cause of this effect is the reduced volumes of news given by the weighting process.

In table B.7 we show the statistics for quarterly cumulated news sentiment. The effects of a longer cumulation period are visible as the news volumes in each period are increased and this is reflected in the indicators increased mean and standard deviation.

In table B.8 we show the statistics for quarterly cumulated fresh news sentiment. As for monthly cumulated fresh news sentiment, the reduced means and standard deviations, compared to the previous table, show the effect of freshness weighting of news pieces. At the same time the effects of a longer cumulation period are visible as the news volumes in each period are increased and this is reflected in the indicators increased mean and standard deviation.

In tables B.9 and B.10 we show the statistics for weekly cumulated total and fresh news sentiment. The effects of a shorter cumulation period are visible as the news volumes in each period are reduced and this is reflected in the indicators reduced mean and standard deviation.

4. Pre-opening news portfolio selection

4.1. The portfolio selection

To inspect the capabilities of the news indicators for asset allocation: we exploit the information brought by the news in a long-short fully invested portfolio strategy. Long and short positions on the market index are used as a reference benchmark.

Each one of the tested strategies takes into account only one news indicator at a time and is fully invested unless there is no news in the considered category group for any of the considered stocks in the considered cumulation period. More than 300 strategies are analyzed.

The strategies are based on a naive beauty contest ([Key36]). Each one of the available stocks is weighted in the portfolio according to the value of the considered news indicator. The indicator values under a predetermined small threshold level are zeroed to avoid the risk to fully invest the portfolio in stocks with an insignificant news indicator value in periods with very few news pieces. The weights are then normalized to sum to 1 in absolute value to obtain an unleveraged fully invested portfolio if enought sentiment is present.

$$\forall\, j \in \mathbb{S}, \forall\, d \ \in \ \{2005/01/01, \dots, 2018/05/31\}$$

$$W_j = \begin{cases} \tau_j, & \text{if } |\tau_j| > \xi \\ 0, & \text{otherwise} \end{cases}$$

$$P_j = \begin{cases} \dfrac{W_j}{\sum\limits_{k\,\in\,\mathbb{S}} |W_k|}, & \text{if } \sum\limits_{k\,\in\,\mathbb{S}} |W_k| > 0 \\ 0, & \text{otherwise} \end{cases}$$

where:
$\mathbb{S}$ is the set of available stocks,
τ_j is the news indicator value for the stock j at day d,
ξ is a small threshold level, equal to 0.05,
W_j is the pure weight for the stock j at day d,

and P_j is the normailzed weight for the stock j in the porfolio at day d.

The obtained portfolio strategies are then evaluated separately for intraday trading on the open to close and close to next day open time frames. To evaluate the strategies the considered indicators are the final wealth, the standard deviation, the skewness, and the kurtosis of the portfolio log-returns. The final log-wealth of a strategy has been found more informative than the linear correlation of the sentiment with the log-returns. Due to the low signal to noise ratio, sentiment indicators not significantly correlated with returns can anyway lead to profitable portfolio strategies.

The formulas used to calculate the final log-wealth, the standard deviation, the skewness, third standardized moment, and the zero centred kurtosis, zero centred fourth standardized moment, of the portfolio strategies are respectively:

$$X = \sum_{j \in \mathbb{S}} P_j R_j$$

$$\pi = nE[X]$$

$$\sigma = \sqrt{E[(X - E[X])^2]}$$

$$\tilde{\mu}_3 = \frac{E[(X - E[X])^3]}{E[(X - E[X])^2]^{3/2}}$$

$$\tilde{\mu}_4 = \frac{E[(X - E[X])^4]}{E[(X - E[X])^2]^2} - 3$$

where:
R_j is the log-return of the j-th stock.
X is the log-return of the portfolio strategy.
n is the number of observations.
π is the final log-wealth of the portfolio strategy.
σ is the standard deviation of the portfolio strategy.
$\tilde{\mu}_3$ is the skewness of the portfolio strategy.
$\tilde{\mu}_4$ is the zero centred kurtosis of the portfolio strategy.

In the analysis, the EURO STOXX 50 market index is considered as a reference benchmark for comparison.

No transaction cost is evaluated at the moment. In intraday trading transaction costs can consistently decrease the profitability of some news based strategies and also make others unprofitable. The reported results can be anyway used to evaluate market entering and exiting strategies in longer horizon strategies.

4.2. The empirical analysis

The evaluation of the generated portfolio strategies shows interesting results. Many of the considered category groups are characterized by patterns of profitable strategies. Some category group is characterized by a very limited number of news pieces, too few to clearly evaluate the strategies under this framework. Some other category groups show a time-varying profitable pattern. Our analysis is focused on stable and profitable strategies. Other results are anyway reported. The stable and profitable strategies show a restricted number of patterns. We observe that news freshness is not an important parameter for most of the examined cases. The RavenPack taxonomy, defining the category groups, instead has a central role in the profitability of the portfolio strategies.

4.2.1. Sell on news volumes

One of the main patterns follows the well known proverbial phrases "No news is good news", and "Sell on the news". The related profitable strategy is to sell the stocks at opening weighting the short portfolio using the volume of news and buying back at closing time. The pattern shows good profitability for many category groups and often a weaker reversal effect overnight. The category groups which follow this pattern are: Acquisition Merger, Analyst Ratings, Assets, Earning, Equity Actions, Labor Issues, Partnership, and Stock Prices.

Table 4.1.: This table shows the final ex-post log-wealth and log-returns standard deviation obtained with each one of the strategies derived from the category groups for total news volume. The values for EURO STOXX 50 market index are also reported at the end of the table as a reference benchmark.

	Log-Wealth		Standard Deviation	
Category Group	**Open-Close**	**Overnight**	**Open-Close**	**Overnight**
Acquisitions-mergers	**-2.709**	**1.901**	0.014	0.007
Analyst-ratings	**-1.78**	**1.43**	0.013	0.008
Assets	**-1.271**	**1.008**	0.012	0.006
Credit	-0.072	0.189	0.008	0.004
Credit-ratings	-0.337	-0.171	0.005	0.005
Dividends	-0.161	0.397	0.009	0.005
Earnings	**-1.157**	**1.433**	0.015	0.008
Equity-actions	**-1.37**	**1.211**	0.014	0.007
Labor-issues	**-2.538**	**0.957**	0.014	0.008
Legal	-0.587	0.06	0.008	0.004
Marketing	0.247	0.051	0.005	0.002
Partnerships	**-1.598**	**0.864**	0.009	0.005
Price-targets	-0.432	**1.09**	0.011	0.007

Products-services	0.468	**1.735**	0.014	0.008
Revenues	-0.359	**1.006**	0.014	0.007
Stock-prices	**-1.751**	**1.763**	0.016	0.009
Technical-analysis	0.36	0.768	0.01	0.006
Market Index	-0.322	0.476	0.01	0.003

The profitability of each strategy is not only related to the average return over the period, but also to the mean-variance trade-off of each return time series and to the stability of the effects over time.

- The **Acquisition Merger** sell on news volume strategy is mainly driven by positive sentiment and was particularly strong during the 2008 financial crisis.

- The **Analysit Ratings** strategy, also mainly driven by open to close negative sentiment, shows an improvement in performance starting from the 2008 financial crisis. The performance is lower from 2014.

- The **Assets** strategy shows good profitability until the end of 2015, and no reversal effect after the second half of 2013.

- The **Earnings** strategy shows on the open to close time frame a highly volatile pattern and violations to the trend especially during the 2009 financial crisis, while the reversal overnight is more stable and consistent.

- The **Equity Actions** strategy shows good profitability before 2009 on the open to close, and a conspicuous overnight reversal from 2009 up to the end of 2015.

- The **Labor-issues** strategy (see figure 4.1) shows good profitability with time-varying performances on the open to close and a limited reversal overnight. The strategy reports peaks of profitability during the crisis in 2008-2009 and 2015-1016, while negative performances are present after the crisis, especially in 2009 during recovery.

- The **Partnership** strategy shows profitability, especially at the end of 2008.

- The **Price-targets** strategy shows periods of profitability for overnight reversal, while for the open to close consistent losses are present in 2011 and 2012.

- The **Product-services** and **Revenues** strategies show profitability for overnight reversal, but a different time-varying pattern is present on the open to close characterized by volatility with periods of growth and periods of losses.

- The **Stock Prices** strategy shows reversal profitability on the overnight interval but high volatility on the open to close.

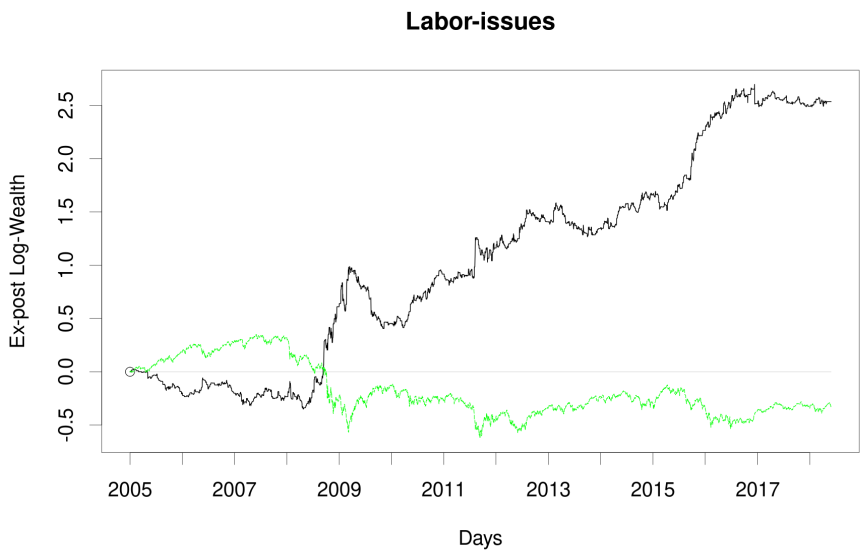

Figure 4.1.: Plot of the sell on 4 hours pre-opening news volumes intraday open to close strategy for the category group "Labor issues" in black, and the open to close reference EURO STOXX 50 market index strategy in green.

Table 4.2.: This table shows the skewness and kurtosis obtained with each one of the strategies derived from the category groups for total news volume. The values for EURO STOXX 50 market index are also reported at the end of the table as a reference benchmark.

	Skewness		**Kurtosis**	
Category Group	**Open-Close**	**Overnight**	**Open-Close**	**Overnight**
Acquisitions-mergers	-2.368	1.618	38.8	41.3
Analyst-ratings	-0.623	-3.897	14.6	180.1
Assets	-3.387	-1.027	65	113.9
Credit	-1.814	1.897	137.5	176.2
Credit-ratings	-3.691	-32.997	115.9	1598.8
Dividends	-1.083	-17.591	47	928.3
Earnings	0.09	-0.128	24.5	46.5
Equity-actions	-0.287	-1.826	33.2	84.4
Labor-issues	-0.878	-0.324	38.2	67.6
Legal	0.597	-2.579	86.4	83.7

Marketing	2.126	2.853	133.2	185.5
Partnerships	-2.797	1.733	42.9	92.3
Price-targets	0.019	-0.895	48	77.1
Products-services	-0.334	-1.131	10.9	35.9
Revenues	-0.019	0.594	14.6	27.8
Stock-prices	-0.51	0.537	26.3	38
Technical-analysis	0.581	-3.616	36.2	80.8
Market Index	-0.233	-0.373	11.1	6.3

4.2.2. Buy according to news sentiment

Another main pattern trades the stocks on the open to close intraday interval weighting the long-short portfolio using the total sentiment of news. The pattern overnight, on the close to next opening interval, shows weak profitability underperforming the market index resembling a weak reversal effect, with some exceptions. The category groups which follow this pattern are: Analyst Ratings, Dividends, Price Targets, Stock Prices, and Technical Analysis.

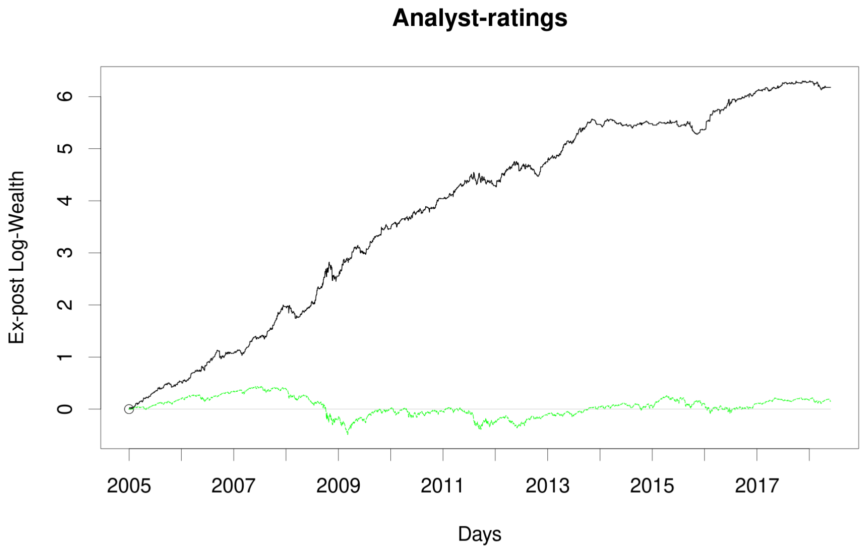

Figure 4.2.: Plot of the buy on 4 hours pre-opening news sentiment open to next opening strategy for the category group "Analyst Ratings" in black, and the open to next opening reference EURO STOXX 50 market index strategy in green.

- The **Analyst Ratings** buy on news sentiment strategy (see 4.2) shows great profitability until 2014 on the open to close time frame, while weak profitability is present for the close to next opening. The open to close profit is driven by negative sentiment, which shows reversal overnight possibly to overreaction, while positive sentiment shows continuation overnight, achieving a similar performance on the open to next opening time frame. The reason for this peculiar pattern could reside in the limited information leaking in the analysts highly regulated sector, as reported in other studies, and the quick reaction to news of the market players.

- The **Dividends** strategy shows periods of profitability on the open to close in 2008 and 2013.

- The **Labor-issues** strategy shows profitability on the open to close up to the end of 2015, and limited reversal with high volatility overnight.

- The **Price Targets** strategy shows profitability on the open to close up to the end of 2012, characterized by peaks at the beginning of 2009 and of 2012, while no reversal is present overnight.

- The **Stock Prices** strategy shows profitability on the open to close but with high volatility, while a strong reversal is present overnight between 2014 and 2016.

- The **Technical Analysis** strategy shows profitability on the open to close up to the end of 2015, while a strong reversal is present overnight, especially from 2011 up to the end of 2016.

4.2.3. Sell according to news sentiment

The last main pattern trades the stocks on the open to close intraday interval weighting the long-short portfolio using the opposite of the news total sentiment. The pattern also shows a reversal effect on the close to the next opening interval, with one exception. The category groups which follow this pattern are: Acquisitions Mergers, Earnings, Partnerships, and Revenues.

Table 4.3.: This table shows the final ex-post log-wealth and log-returns standard deviation obtained with each one of the strategies derived from the category groups for total news sentiment. The values for EURO STOXX 50 market index are also reported at the end of the table as a reference benchmark.

	Log-Wealth		**Standard Deviation**	
Category Group	**Open-Close**	**Overnight**	**Open-Close**	**Overnight**
Acquisitions-mergers	**-2.252**	**1.684**	0.014	0.007
Analyst-ratings	**5.847**	0.334	0.012	0.007
Assets	-0.317	0.118	0.009	0.005

Credit	0.515	0.098	0.005	0.003
Credit-ratings	0.351	-0.128	0.005	0.005
Dividends	**0.573**	0.202	0.007	0.003
Earnings	**-2.206**	0.295	0.014	0.007
Equity-actions	-0.544	-0.366	0.014	0.007
Labor-issues	**1.644**	0.048	0.013	0.007
Legal	0.747	0.042	0.008	0.004
Marketing	0.163	0.019	0.002	0.001
Partnerships	**-1.357**	0.683	0.009	0.005
Price-targets	**2.342**	0.499	0.01	0.006
Products-services	0.747	0.832	0.013	0.008
Revenues	**-1.773**	**0.734**	0.013	0.006
Stock-prices	**1.478**	0.228	0.016	0.008
Technical-analysis	**1.274**	**-1.011**	0.01	0.006
Market Index	-0.322	0.476	0.01	0.003

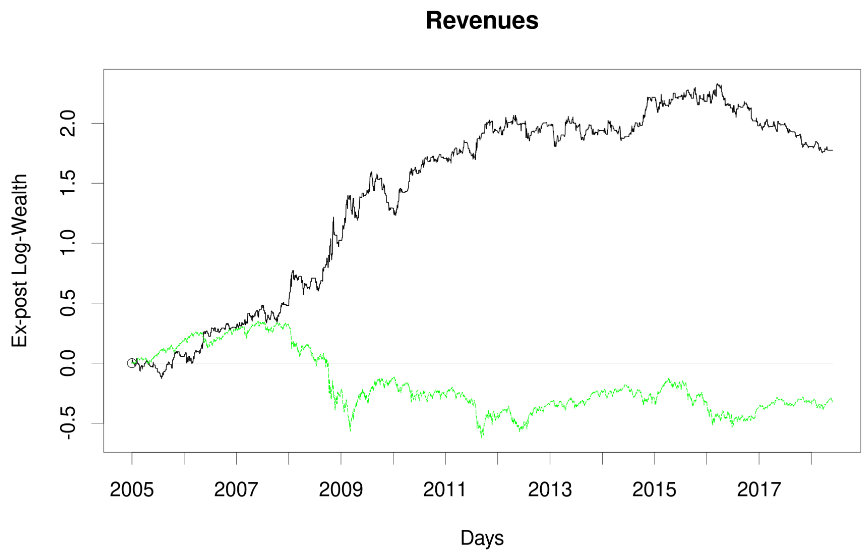

Figure 4.3.: Plot of the sell on 4 hours pre-opening news sentiment intraday open to close strategy for the category group "Revenues" in black, and the open to close reference EURO STOXX 50 market index strategy in green.

- The **Acquisition-Mergers** strategy shows time-varying profitability on the open to close and a less volatile reversal overnight. A period of extreme profitability and a limited reversal is present during the 2008 financial crisis.

- The **Earnings** sell on news sentiment strategy shows good profitability, especially for fresh news, on the open to close interval before 2013, while overnight a time-varying pattern is present showing reversal before 2008 and after 2013.

- The **Partnerships** strategy shows time-varying profitability on the open to close interval, while overnight a weak reversal is present. A peak of profitability and a limited reversal is present during the 2008 financial crisis.

- The **Revenues** strategy (see 4.3) shows profitability on the open to close interval before 2016, especially before 2012 when the profitability suddenly decreases, while overnight a weak reversal is present, especially in 2008.

Table 4.4.: This table shows the skewness and kurtosis obtained with each one of the strategies derived from the category groups for total news sentiment. The values for EURO STOXX 50 market index are also reported at the end of the table as a reference benchmark.

	Skewness		**Kurtosis**	
Category Group	**Open-Close**	**Overnight**	**Open-Close**	**Overnight**
Acquisitions-mergers	-2.356	1.436	38.6	42.3
Analyst-ratings	0.59	-5.887	17.5	131.9
Assets	-2.613	2.029	105.5	119
Credit	5.862	-0.847	264	279.7
Credit-ratings	4.94	-24.104	134	1670.1
Dividends	3.984	-0.076	150.2	185.4
Earnings	-0.883	-0.24	28	53.2
Equity-actions	-1.426	-3.63	44.3	94.1
Labor-issues	-1.008	-0.81	37.9	73.4
Legal	0.886	3.396	85.7	84.5
Marketing	8.76	31.204	395.7	1963.6
Partnerships	-1.779	1.414	43.4	93.4
Price-targets	2.895	-0.826	60.2	53
Products-services	-0.297	-1.634	11.7	39.5
Revenues	-0.593	-0.032	18.9	21.6
Stock-prices	0.429	-0.397	29.2	45.5
Technical-analysis	1.655	0.09	52.4	95.7
Market Index	-0.233	-0.373	11.1	6.3

4.2.4. Other candidates

Some other category groups show profitable patterns but too limited news volumes are present to implement a fully invested strategy and thus a different benchmark and methodology should be taken into account to clearly analyze the results. The indicators showing to be good candidates for profitable strategies also if with a limited volume of news are:

- **Crime** negative news sentiment.
- **Indexes** fresh positive sentiment overnight.
- **Insider Trading** total sentiment open to close.
- **Marketing** total sentiment open to next opening.
- **Regulatory** negative news sentiment.
- **War Conflict** positive news sentiment open to close.

4.3. Further improvements

Many improvements may be added in future researches:

- A calibration step could be performed. For each category group the optimal lookback window, holding horizon, and also minimum sentiment threshold could be determined.
- The sentiment effect may be better modelled by determining an appropriate exponent for the sentiment value to vary the allocation concentration among stocks in the beauty contest, overweighting or underweighting stocks with a strong or weak sentiment.
- The wealth invested in the portfolio could be limited when the total sentiment is under a given threshold, avoiding fully investing in the portfolio when the sentiment is ambiguously near to neutral.
- A strategy that jointly considers the more profitable category groups correctly discounting the cross effects could also be considered.

5. Portfolio selection with news

This section describes a model aimed at augmenting a baseline portfolio selection model with the use of news. The model is aimed at calculating the optimal proportion of assets to hold for non-satiable risk-averse investors. The models is based on portfolio selection theory that developed from the work, between the others, of Tobin and Markowitz (see [Mar59], [Tob58], and [Tob65]).

News sentiment effects are reported to affect prices up to two weeks after the announcement. If the portfolio turnover is partial the reallocation process can anyway take place on a daily basis and, due to the low turnover, exploit longer-lasting trends, for example on a 10 day period.

After having taken into account the peculiarities of news sentiment data, reported in the previous chapter, we have identified a group of cumulated news sentiment indicators that are reputed suitable for this kind of analysis. The sentiment indicators differ by the considered news event category group, but not all the indicators are suitable for the analysis.

Many of the category groups are not suitable for the analysis because too few news pieces are present in the considered period. Especially news events category groups with less than one news piece per quarter are reputed not suitable. Some category group is also characterized only by neutral sentiment. Category groups characterized only by neutral news sentiment are not suitable for this model because news volumes are not directly taken into account.

5.1. Risk-return optimization

To estimate the optional proportion of assets to hold for each investor a criterion to be maximized is needed. In this work, three different criteria have been considered and are hereafter reported.

5.1.1. Sharpe Ratio maximization

To inspect the capabilities of the news indicators to improve the portfolio performance, the idea is to start from a long-only fully invested baseline portfolio strategy, used also as a reference benchmark, and to try to enhance it using the information inferred from the news indicators.

The portfolio weights for the baseline strategy are calculated maximizing the Sharpe ratio of the portfolio log-returns (see [Sha66] and [Sha94]).

The portfolio log-return X is given by the weighted sum of the stocks' log-return R_j:

$$X = \sum_{j \in \mathbb{S}} P_j R_j$$

where:
$\mathbb{S}$ is the set of available stocks,
P_j is the weight of the j-th stock,
R_j is the log-return of the j-th stock,
and X is the log-return of the portfolio.

The Sharpe Ratio is calculated as the mean of the portfolio log-return minus the risk-free rate of return r_{rf}, divided by the standard deviation of the portfolio log-return:

$$\frac{E\left[X\right] - r_{rf}}{\sqrt{E\left[\left(X - E\left[X\right]\right)^2\right]}}$$

The long-only fully invested constraint implies that the sum of the weights must be equal to 1:

$$\sum_{j \in \mathbb{S}} P_j = 1$$

and that each weight P_j must be greater than 0:

$$P_j \geq 0 \; \forall j \in \mathbb{S}$$

The optimal portfolio is thus found searching for the set of weights P that maximize the portfolio Sharpe ratio, without violating the long-only fully invested constraint:

$$\max_P \left(\frac{E\left[\sum_{j \in \mathbb{S}} P_j R_j\right] - r_{rf}}{\sqrt{E\left[\left(\left(\sum_{j \in \mathbb{S}} P_j R_j\right) - E\left[\sum_{j \in \mathbb{S}} P_j R_j\right]\right)^2\right]}} \right)$$

$$\text{s.t.} \sum_{j \in \mathbb{S}} P_j = 1 \;, \; P_j \geq 0 \; \forall j \in \mathbb{S}$$

where:
P is a vector of weights,
and r_{rf} is the risk-free rate of return.

5.1.2. Second-order Stochastic Dominance

Second-order Stochastic Dominance (SSD) is a different criterion for portfolio weights optimization given the stock log-return panel (see [HR69], [HL69], [Lev92], and [Lev90]). An SSD dominant portfolio is in line with the portfolio choices of any non-satiable risk-averse investor, regardless of his/her utility function.

Definition: A stock return distribution R_j is then said to dominate another stock return distribution R_k with respect to Second-order Stochastic Dominance, in symbols $R_j \succeq_{SSD} R_k$, if the following inequality holds for every value of x:

$$\int_{-\infty}^{x} F_{R_j}(t)\,dt \leq \int_{-\infty}^{x} F_{R_k}(t)\,dt$$

where:
F_{R_j} is the cumulative distribution function of the log-returns for stock j and
F_{R_k} is the cumulative distribution function of the log-returns for stock k.

The search for the dominant portfolio requires a multi-objective optimization approach, obtained minimizing the conditional value at risk (CVaR), or expected shortfall (ES), for every value of α:

$$CVaR_\alpha(X) = \frac{1}{\alpha}\int_{0}^{\alpha} VaR_\gamma(X)\,d\gamma$$

where $VaR_\gamma(X)$ is the γ value at risk of the random variable X.

The value at risk is defined such that the probability of a loss greater than VaR is γ. Specularly the probability of a loss smaller than VaR is $1-\gamma$. The VaR_γ of X correspond to the opposite in sign of γ-th quantile in the X log-return distribution:

$$VaR_\gamma(X) = -F_X^{-1}(\gamma)$$

where F_X^{-1} is the inverse of the cumulative distribution function of X.

Equivalently the CVaR can be computed as the expected value of X given that X is less than the γ value at risk of X:

$$CVaR_\alpha(X) = -E[X|X < -VaR_\alpha(X)]$$

Stochastic Dominance is a criterion based on ordering. A return distribution can dominate another if all the constraints are satisfied, or the ordering can not be established if the constraints are fulfilled for some but not all the values of α.

The optimization problem, considering daily stocks log-returns, reduces to the discrete space. The return distribution panel is assumed to be given by a finite number of samples extracted from the more recent history of each stock. If N equiprobable concurrent observations are considered for each stock, α can assume a finite number of values. In detail α can assume values in $\{n/N \mid 1 \leq n \leq N\}$ and if each stock return R_j is sorted in ascending order the CVaR can be easily computed as the opposite of the expected value of the first $\alpha * N = n$ samples:

$$CVaR_{\frac{n}{N}} = -\sum_{i=1}^{n} \frac{R_{i:N}}{n}$$

where $R_{i:N}$ is the i-th observation in the increasing ordered of return R.

The multi-objective optimization problem of finding the stock weights of the optimal portfolio can thus be formulated as a minimization problem:

$$\mathbb{Z} = \left\{ \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N}{N} \right\}$$

$$\min_{P} \left(\left\{ CVaR_{\alpha} \left(\sum_{j \in \mathbb{S}} P_j R_j \right) \mid \alpha \in \mathbb{Z} \right\} \right)$$

$$\text{s.t. } \sum_{j \in \mathbb{S}} P_j = 1 \ , \ P_j \geq 0 \ \forall j \in \mathbb{S}$$

where $\mathbb{Z}$ is the set of values α available in the discrete space.

5.1.3. Enhanced indexation and index dominance

The SSD is a computationally intensive task and sometimes the criterion is considered too restrictive. An alternative method, optimizing the SSD in respect to a benchmark, have therefore been developed. In enhanced indexation, the stock market index is taken as a reference as an ex-ante optimal allocation portfolio weighting and thus CVaR structure. This limitation reduces the search problem to a specific minimum dominating the reference index. But this could be considered only one among any other reference portfolio that could be used, for instance, the portfolio resulting from the baseline strategy hereafter presented could be used as a reference for the improved strategy.

The multi objective optimization is solved trying to minimize for the worst α, that is the one with the maximum "negative" difference between the portfolio conditional value at risk and the reference benchmark:

$$\min_{P} \left(min \left(\left\{ CVaR_{\alpha} \left(\sum_{j \in \mathbb{S}} P_j R_j \right) - CVaR_{\alpha}(I) \mid \alpha \in \mathbb{Z} \right\} \right) \right)$$

$$\text{s.t. } \sum_{j \in \mathbb{S}} P_j = 1 \ , \ P_j \geq 0 \ \forall j \in \mathbb{S}$$

where I is the reference index log-return distribution.

In the optimization, the long-only fully invested constraint, already presented for Sharpe Ratio, is maintained.

The optimization problem is convex and can be solved with a linear programming model (see [RMZ11]).

5.1.4. Scaled SSD

Second-order Stochastic Dominance sometimes is still a too restrictive criterion. Sometimes a dominant portfolio can not be found or do not exist even for the second-order, because not all the constraints can be satisfied simultaneously. For this reason, more relaxed versions of stochastic dominance have been proposed, other choices can regard for instance Third-order Stochastic Dominance controlling for the mean, Almost First-order and Almost Second-order Stochastic Dominance, Zero-order ϵ Stochastic Dominance (see [BCST17]), or scaled versions of the SSD.

In this work, we follow the approach of Scaled SSD proposed by [FMRZ11] and applied in [RMZ11]. In the scaled version the scaled tails, CVaR considered at the confidence levels $\frac{i}{S}$ with $i \in \{1 \dots N\}$, are considered instead of the original CVaR.

The modified constrained optimization then becomes:

$$\min_{P} \left(min \left(\left\{ \alpha \left(CVaR_{\alpha} \left(\sum_{j \in \mathbb{S}} P_j R_j \right) - CVaR_{\alpha}(I) \right) \mid \alpha \in \mathbb{Z} \right\} \right) \right)$$

$$\text{s.t.} \sum_{j \in \mathbb{S}} P_j = 1 \ , \ P_j \geq 0 \ \forall j \in \mathbb{S}$$

The under-weighting of extreme events in the scaled version allows to partially avoid the problem of non-eludible fat-tails in the return distributions (stock weight could be limited in an SSD optimized portfolio because it has just one very negative outcome in its history, regardless of all other outcomes).

The weights applied to each value of the CVaR are thus equal to the considered parameter α, this operation also cancels out (simplifies) in the algorithm implementation as it is the inverse operation of the integral mean used to compute the Expected ShortFall.

5.2. The optimizer

The Sharpe ratio maximization and Second-order Stochastic Dominance are performed applying the SUB PLEX algorithm (see [Row90]), a variation of the Nelder-

Mead derivative-free algorithm for convex optimization. The derivative-free algorithm is reputed suitable for optimization because the optimization process consumes only a fraction of the computational time of the whole process, involving beyond that also multiple principal component analysis and linear regressions. The constraint, implying that the sum of the portfolio weights has to be equal to 1, is transformed into a normalization of the weights inside the fitness function, due to the lack of support to optimization constraints under SUB PLEX. The solution using SUB PLEX with weight normalization was tested and resulted to be equal to a constrained optimization performed using the slower, derivative-free version, of the algorithm COBYLA.

5.2.1. Dynamical market models

According to [UM16], [CDK12] and [KSL11], to take into account the dynamic market conditions and news effects on the log-returns, the effects are analyzed separately for sub-periods using a look-back window.

In this work many different rolling windows are present, mainly subdivied into three steps: the identification and regression of the market common trends, the identification and regression of the news principal components, and the optimization of the portfolio weights through the three optimization criteria considered. For the sake of simplicity, all the rolling windows have been considered equal and with a length of 3 years. The 3-year window is in line with the previously reported analysis of market predictability for the Adaptive Market Hypothesis (see [Lo05]). The results are then reported starting from 2008 when a rolling window is needed because the previous period is used for pre-calculations only.

5.3. Factors from Principal Components Analysis

Principal Component Analysis (PCA) is a common tool in statistics. The PCA and Principal Axis Factorization (PAF), one of its main extensions, are tools able to extract the main co-moving factors from a set.

The number of factors in a regression affects the out of sample performance. A growing number of independent factors introduce more and more noise into the regression and increase the probability to exploit spurious correlations.

The figure 5.1 clearly shows that the variance of a dependent variable can be easily explained by spurious factors that have nothing in common and are not related to the dependent variable. In the figure random normally distributed variables are considered. The correlation between the variables “in the large numbers” should theoretically converge to 0, but, in the finite case, as the ratio between the number of observations and the number of factors approaches to 1, on average the linear model erroneously explains part of the variance of the dependent variable using the spurious factors.

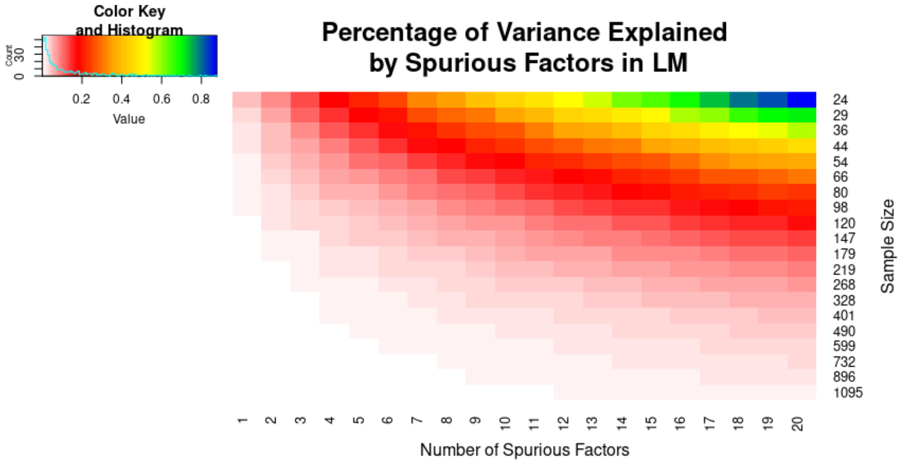

Figure 5.1.: The image shows the average percentage of explained variance by a linear model including only spurious factors. The percentage is influenced by the sample size and the number of factor used.

As shown by [PPNK05] and [KPN07] the ratio between the number of assets and the number of observations must be small, less than 1, and possibly near to 0 to obtain a good approximation of the portfolio risk-reward measures, but this is in contrast with the time-changing dynamical market model. One of the possible solutions is to reduce the randomness approximating the stock returns by applying the k-fund separation model. In the market-plus-sectors model, the covariance is fixed and differ only for assets belonging to the same market sector. The principal components represent the few factors with the highest return variability. The original stock returns are then replaced with a linear combination of the principal components with significant variability, while the others are summarized by an error.

5.3.1. Market common trends

The market common trends are extracted applying the principal component analysis cross-sectionally on the correlation matrix of the stocks returns and these coincide with the first few principal components returned (see [RMF$^+$07]). The correlation matrix of the log-returns is a positive semi-definite matrix and thus its eigenvectors ordered by eigenvalue represents the ordered set of principal components. The principal components are computed using the Singular Value Decomposition algorithm that is numerically more accurate than directly computing the Eigen-decomposition on the variance-covariance matrix. In each period, the stocks with more than 2% of missing values are removed from the analysis and the remaining missing values are set to 0 as those are supposed to be relative to non-trading days and thus the stock price did not change. The log-returns are centred and scaled to compensate for the differences in the log-return distributions between different stocks, and especially for differences in volatility. The first m components with the largest eigenvalues, and thus explained variance, are interpreted as the common trends.

The problem of finding the m orthogonal vectors V_i can be equivalently formulated as the minimization of the representation error given by the mean square distance between the original data and the back-projection into the original space of the data projected into the lower dimensional space of m dimensions:

$$N_j = \frac{R_j - E\left[R_j\right]}{\sqrt{E\left[\left(R_j - E\left[R_j\right]\right)^2\right]}}$$

$$\min_V \; E\left[\left\| N_j - \sum_{i=1}^{m} \langle V_i, N \rangle V_i \right\|^2\right], \; m < n$$

where:
N_j is the standardized log-return of the j-th stock,
m is the number of principal components considered,
n is the cardinality of $\mathbb{S}$, number of available stocks,
V_i is the i-th ordered eigenvector of the variance-covariance matrix of N,
V_{ij} is the weight of the j-th stock for the i-th common trend,

The market common trends can then be reconstructed as:

$$M_i = \langle V_i, R_j \rangle = \sum_{j=1}^{n} V_{ij} R_j$$

where M_i is the i-th common trend.

The value chosen for the number m of market common trends used is set to 6. The ratio between the number of observations and the number of factors (as recognized by [PPNK05] and [KPN07]) have to be strictly greater than 1 and the results improve for increasing values. The choice has been made keeping in mind the reported results and also after taking into account the standard deviation explained by the principal components. Furthermore in literature is often considered a number of factors between 5% and 10% of the number of stocks.

5.3.2. Cross-sectional news factors

Principal factors regarding the news cumulated sentiment indicators are extracted, as previously done for market common trends, applying the principal component analysis cross-sectionally to the standardized news sentiment indicators separately for each news category group. The considered category groups are those reputed possibly significant after the news indicator analysis developed in chapter 3.1.3. The selected category groups are those with on average at least one news piece per stock each month and with sentiment values significantly different from 0. The principal components are extracted considering again 3 years of historical data at each step and moving forward the window at each portfolio reallocation step.

$$\forall\, c \in \mathbb{C}, \forall\, p \in \mathbb{P} = \{30, 90\}$$

$$\chi_j = \frac{\tau_j - E\left[\tau_j\right]}{\sqrt{E\left[\left(\tau_j - E\left[\tau_j\right]\right)^2\right]}}$$

$$\min_W\ E\left[\left\|\chi_j - \sum_{i=1}^{l} \langle W_i, \chi\rangle W_i\right\|^2\right],\ l < n$$

where:
χ_j is the standardized cumulated news sentiment of the j-th stock for a given category group c and cumulation period p
l is the number of principal components considered,
n is the cardinality of $\mathbb{S}$, number of available stocks,
W_k is the k-th ordered eigenvector of the variance-covariance matrix of χ
W_{kj} is the weight of the j-th news indicator for the k-th news principal component.

The news principal components can then be reconstructed as:

$$\nu_k = \langle W_k, \chi\rangle = \sum_{j=1}^{n} W_{kj}\chi_j$$

where ν_k is the k-th news principal component for the considered category group c.

The principal components are computed using the Singular Value Decomposition algorithm that is numerically more accurate. The news indicators are standardized to compensate for the different news media coverage of different stocks and some of the potential news sentiment bias affecting particular stocks.

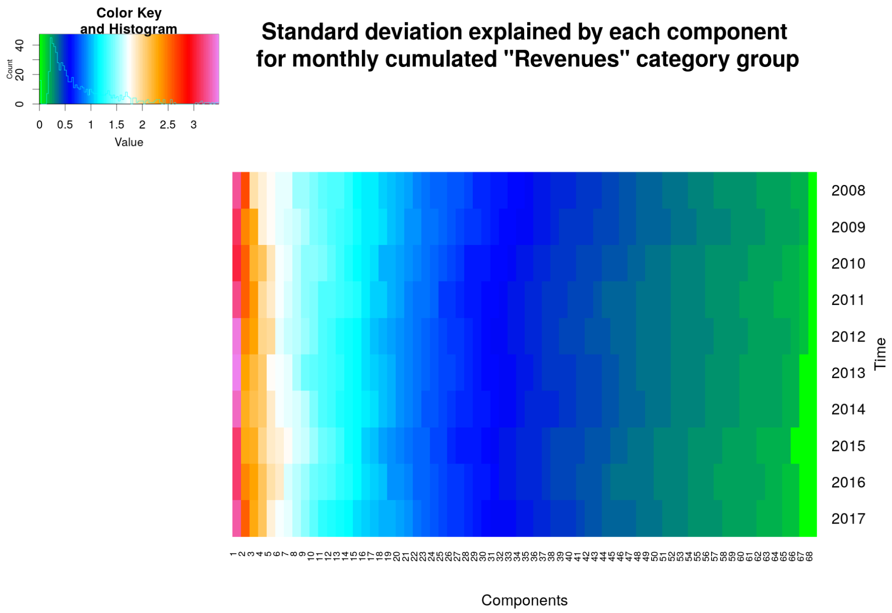

Figure 5.2.: Plot of the standard deviation explained by each one of the principal components for the monthly cumulated news sentiment of the category group “Revenues”

The value chosen for the number l of news principal components characterizing each category group was set to 2. The choice is made, as previously reported, with the aim to keep an observations to factors ratio of at least 100 when all factors are considered together. The choice has been made also after taking into account the standard deviation explained by the news principal components. Especially checking that the eigenvalues relative to the firsts considered principal components are strictly larger than 1. Larger eigenvalues represent a larger number of correlated factors, and principal components summarize their correlated movement. Few large eigenvalues strictly greater than 1 make the application of the PCA useful in respect to the original time series.

As shown in figure 5.2 the standard deviation explained by the principal components is quite stable through time. The components with eigenvalues greater than 1, in the considered period and for the selected category group “Revenues”, are approximately 20. The first 2 or 3 components, on the left in the figure, explain most of the variance, then the explained standard deviation fade to 0 with increasing speed, with many of the last components that explain almost nothing, having eigenvalues near to 0.

The factors extracted from the PCA are applied in the following section to enhance and test two different models.

5.4. Regression model

The approximation of the stock returns is justified by the k-fund separation model according to [Ros76] and [Ros78] and here it is used as baseline model. To adjust the model to consider the news effects the residual of the baseline model are regressed on the news principal components.

The two models, the baseline model and the news enhanced model, are aimed at improving the stability and accuracy of the risk measures applied for portfolio construction and to denoise the stock log-returns keeping only well-established correlations and exploitable volatility. In the first model, the effect given by market common trends is evaluated (see 5.4.1). In the second model is added, to the baseline model, the effect of news principal components on the baseline model residual, and the resulting model is evaluated. The aim of this model is to understand if the news effect is able to improve the quality of the risk measure associated with the volatility signal, compared to the one generated by market common trends (see 5.4.2).

5.4.1. Baseline reference model

The filtered log-returns are generated, cleaned from what is supposed to be market noise, regressing separately the log-returns of each stock on the historical data of the selected market common trends and discarding the residuals.

The regression is defined as:

$$R_j = \beta_{0j} + \sum_{i=1}^{m} \beta_{ij} M_i + \epsilon_j$$

where:
R_j is the log-return of the j-th stock,
M_i is the i-th market common trend,
m is the number of common trends considered in the regression,
the βs are the coefficients obtained regressing the stocks log-returns on the market common trends,
and ϵ_j is the residual error of the model.

The principal components are extracted from the cross-sectional historical data considering a rolling window with a length of 3 years (about 750 observations). The stocks log-returns are regressed on the first m principal components using a rolling window with a length of 3 years.

The choice to use a 3-year rolling window is done after taking into account other different lengths of 2, 5 and 7 years, and it is in line with the choice to maintain an observation to factors ratio of at least 100 when also news factors are considered.

The approximated log-returns are then given by:

$$\widetilde{R}_j = \beta_{0j} + \sum_{i=1}^{m} \beta_{ij} M_i$$

where $\widetilde{R}_j$ is the approximated log-return for the j-th stock.

5.4.2. News enhanced residual model

In this model the stock returns are regressed on the market common trends as done in the baseline model (See 5.4.1). The news factors are computed applying the principal component analysis cross-sectionally separately for each news category group. The residuals of the baseline model are regressed on the news principal components:

$$\epsilon_j = \gamma_{0j} + \sum_{i=1}^{n} \gamma_{ij} \nu_i + \zeta_j$$

where:
ϵ_j is the time series of the residual of the baseline model for the j-th stock,
ν_i is the selected news category group i-th principal factor,
n is the number of principal factors considered in the regression,
the γs are the coefficients obtained regressing the stocks residuals on the news principal factors,
γ_{0j} is equal to 0 as OLS model residuals are zero centered by definition,
and ζ_j is the residual error of the model.

The composition of the fitted part of the two regressions is used as an approximation of the returns:

$$S_j = \widetilde{R}_j + \sum_{i=1}^{n} \gamma_{ij} \nu_i$$

where:
$\widetilde{R}_j$ is the log-return of the baseline model for the j-th stock,
α is the effect ratio parameter balancing the model effect,
γ_{0j} is not reported as it should be equal to 0,
S_j is the approximated log-return for the j-th stock.

5.5. Discussion of results

In this section, the possibility to improve a baseline model with the use of news belonging to relevant predetermined category groups is discussed.

5.5.1. Empirical results without the news

The baseline model is presented with three different possible optimization criteria and evaluated as a reference. The model is then augmented with news coming from selected category groups each one related to a relevant firm-specific event type.

In the baseline model, in line with the k-fund separation model, the original stock time series are approximated regressing the time series on the market common trends obtained from the Principal Component Analysis, according to [PPNK05] and [KPN07]. The stock time series cleaned from noise are then used to estimate optimal portfolio weights using three different risk/reward criteria. The first criterion applied is the Sharpe Ratio optimization (figure 5.3) aimed at evaluating how well the return of an asset compensates the investor for the risk taken.

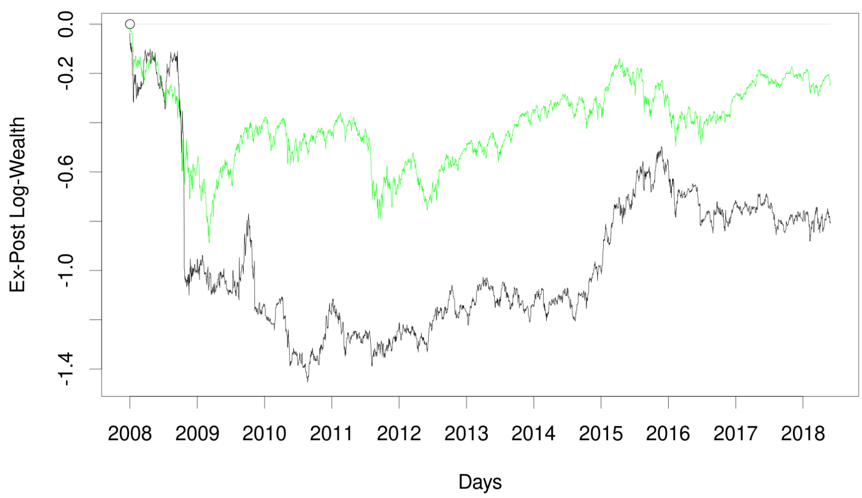

Figure 5.3.: Plot of the strategy based on the Sharpe Ratio criterion, using a 3 years rolling window and 6 market common trends, in black and the EURO STOXX 50 market index in green, as reference for comparison, in the considered period.

The second and the third criteria are respectively based on Second-order Stochastic Dominance (figure 5.4) and its scaled version (figure 5.5). In this framework, we consider that the optimal portfolio must stochastically dominate (with respect to SSD or Scaled SSD) a benchmark. According to the Second-order Stochastic Dominance, the Conditional Value at Risk of the portfolio is compared with that of the benchmark for every considered confidence level. The market index is considered as refenece benchmark for stochastic dominance, and thus the portfolio optimization respect to the market index result in an enhanced indexation if the process is successful.

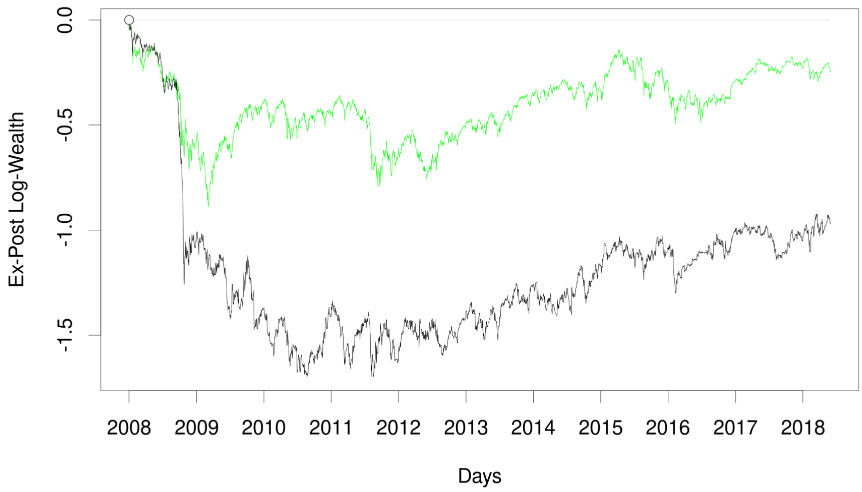

Figure 5.4.: Plot of the strategy based on the Second-order Stochastic Dominance criterion, using a 3 years rolling window and 6 market common trends, in black and the EURO STOXX 50 market index in green, as reference for comparison, in the considered period.

In the scaled version of Second-order Stochastic Dominance, we compare the Conditional Value at Risk of the portfolio and of the benchmark weighted by their confidence level. Under this assumption more extreme tail events result to be underweighted, avoiding potentially undesirable situations in which some stocks could result strongly underweighted due to the presence of few extreme negative outcomes despite the overall positive performance.

According to figures 5.3, 5.4, 5.5, and table 5.1 the three optimization criteria show different performances. Two out of three criteria, Sharpe Ratio and SSD, suffer severe losses during the European financial crisis (2008) compared to the market index. Both strategies underperform the market index, seems to be profitable only starting from 2011, and are characterized by increased volatility compared to the index. The strategy based on Second-order Stochastic Dominance produces worse performance. Counterintuitively, as it should be the most risk-averse, it is also the strategy with the largest drawdown achieved in the period between 2008 and 2011. The strategy based on Sharpe Ratio is profitable starting from 2011, up to the end of 2015, and it clearly outperforms the market index in 2015.

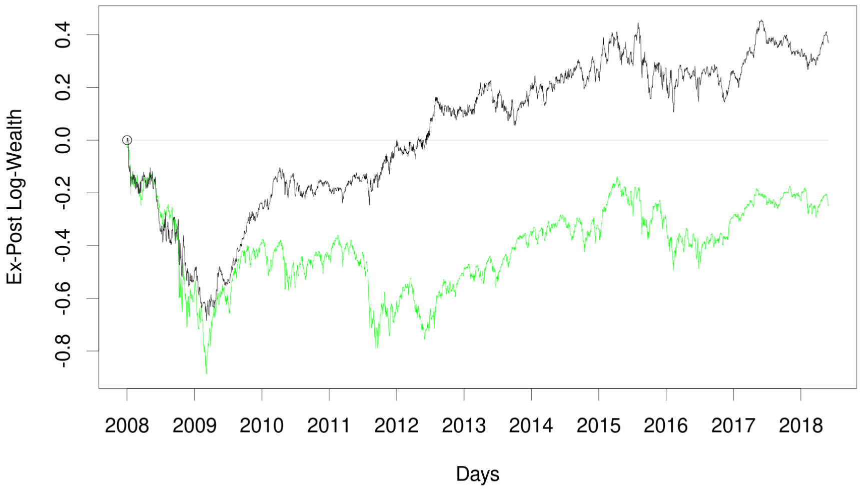

Figure 5.5.: Plot of the strategy based on the scaled version of Second-order Stochastic Dominance criterion, using a 3 years rolling window and 6 market common trends, in black and the EURO STOXX 50 market index in green, as reference for comparison, in the considered period.

The strategy based on the scaled version of Second-order Stochastic Dominance is able to reduce losses during the crisis in 2008 and experience increased gains during the recovery. The profitable period for this strategy starts in 2009, and similarly to the others lasts up to the end of 2015. This strategy is the only one outperforming the market index over the entire period considered, including the crisis in 2008, and it is characterized by volatility slightly lower than the index.

The three strategies are charachterized by very different turnover levels (See figures 5.6, 5.7, 5.8, and table 5.1). The Sharpe Ratio has the lowest turnover of the three, and its turnover level seems strongly regulated by the length of the lookback period considered[1]. The Sharpe Ratio is characterized by just one single risk measure, the standard deviation, and the resulting turnover is lower in respect to the other considered strategies based on stochastic dominance, probably also for this reason. The strategy alternates periods of very low turnover with periods of moderate turnover. The two stochastic dominance strategies, especially the scaled version, are characterized by a higher turnover that could lead to unprofitable strategies due to transaction costs in real situations, if the daily rebalancing horizon constraint is not extended to a longer period[1], such for example at least 5 days.

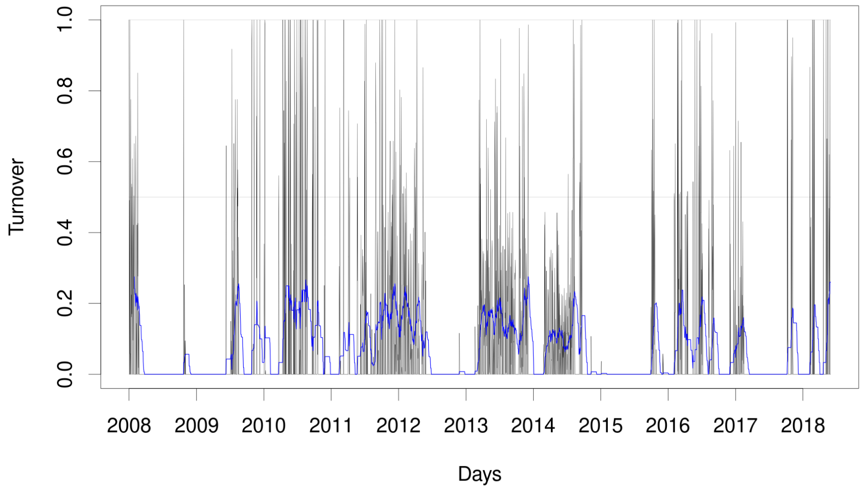

Figure 5.6.: Plot of the Sharpe Ratio strategy turnover in black, and a 30 days turnover average in blue.

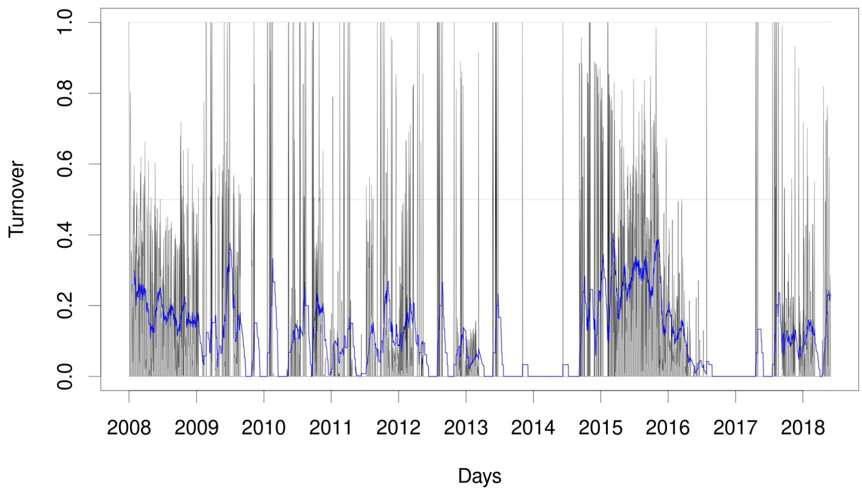

Figure 5.7.: Plot of the Second-order Stochastic Dominance strategy turnover in black, and a 30 days turnover average in blue.

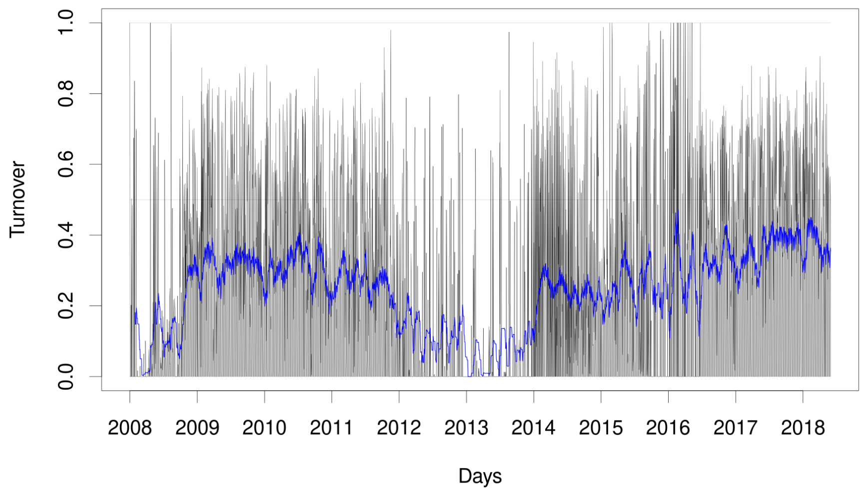

Figure 5.8.: Plot of the scaled Second-order Stochastic Dominance strategy turnover in black, and a 30 days turnover average in blue.

The SSD optimization strategy, as the Sharpe Ratio one, alternates periods of very low turnover with periods of moderate turnover, but the low turnover periods are less and the moderate turnover periods stronger. In the scaled version the turnover is generally higher and seems reduced only in 2008 and in the profitable period of 2013.

Table 5.1.: This table shows the final ex-post log-wealth, returns standard deviation, turnover, and turnover standard deviation for the EURO STOXX 50 Market Index and the 3 considered baseline strategies: Sharpe Ratio optimization, Second-order Stochastic Dominance, and the scaled version of SSD.

	Market Index	**Sharpe Ratio**	**SSD**	**Scaled SSD**
Final Log-Wealth	-0.246	-0.808	-0.97	**0.369**
Standard Deviation	0.013	0.015	0.015	**0.012**
Turnover	-	**0.071**	0.107	0.244
Turnover Std.Dev.	-	**0.195**	0.228	0.281

[1]We tested several in sample/out of sample windows, considering different lookback periods, number of market common trends, and rebalancing period for each strategy in search for profitable strategies. Results show that the PCA increase the portfolio performance, but extra performance is reduced when a rebalancing period longer than daily is applied. The results for Sharpe Ratio show that the lookback period for the mean drives the strategy profitability and turnover. Shorter lookback periods, for instance one year, are more profitable while longer ones reduce the strategy turnover. Results for the Scaled SSD show that longer lookback windows (i.e. of 3 or 4 years) produce more profitable strategies in the long run and show better performances in periods of crisis.

5.5.2. Empirical results with news

The news enhanced models try to improve the reconstruction of the denoised signal regressing the residuals of the baseline model on the firsts principal components of the news cumulated sentiment time series. The principal components of the news sentiment indeed may reflect part of the latent factors of the market. Since different category groups are considered separately, the news factors are not supposed to correctly incorporate all the information regarding the market movements. The market common trends instead are supposed to better reflect the main market movements. Therefore the common trends are discounted at first, and then the news factors are evaluated on the residuals.

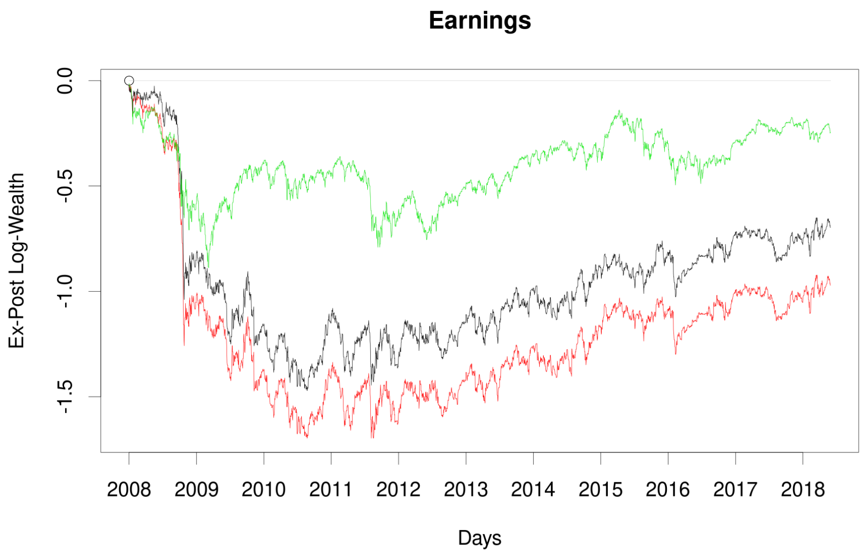

Figure 5.9.: Plot of the strategy based on the Second-order Stochastic Dominance criterion, using a 3 years rolling window and 6 market common trends, enhanced with monthly cumulated "Earnings" total news sentiment in black, the relative baseline reference strategy in red, and the market index in green.

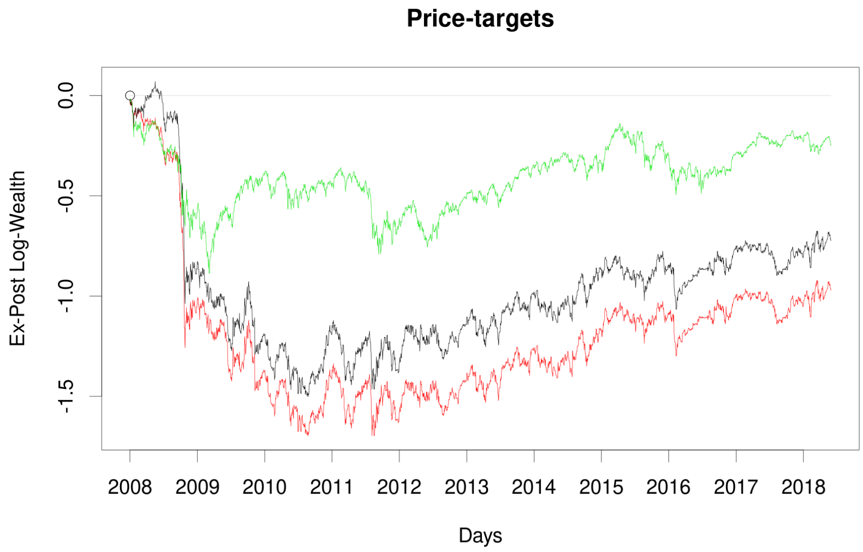

Figure 5.10.: Plot of the strategy based on the Second-order Stochastic Dominance criterion, using a 3 years rolling window and 6 market common trends, enhanced with monthly cumulated "Price-targets" fresh news sentiment in black, the relative baseline reference strategy in red, and the market index in green.

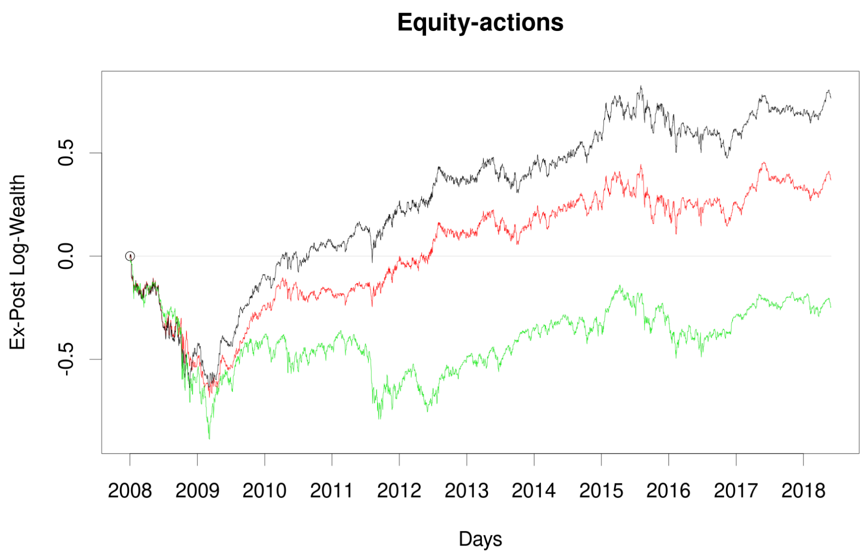

Figure 5.11.: Plot of the strategy based on the scaled version of Second-order Stochastic Dominance criterion, using a 3 years rolling window and 6 market common trends, enhanced with weekly cumulated "Equity-actions" fresh news sentiment in black, the relative baseline reference strategy in red, and the market index in green.

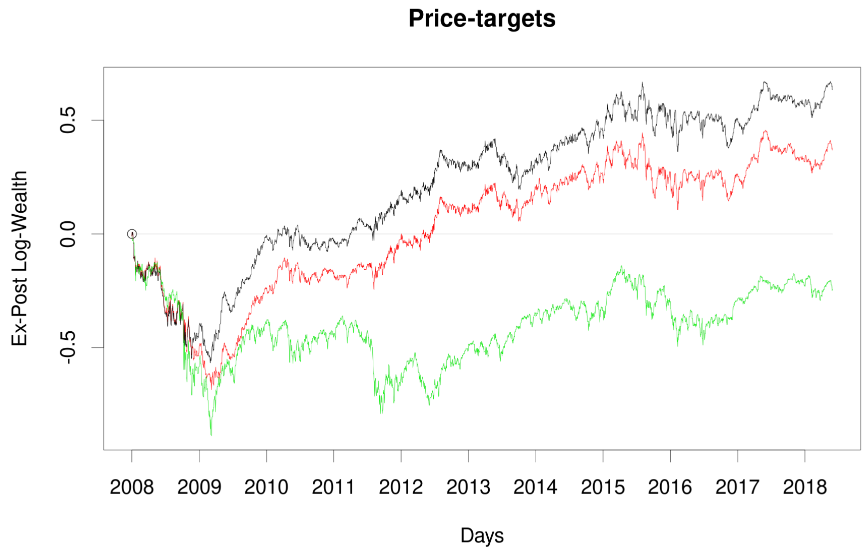

Figure 5.12.: Plot of the strategy based on the scaled version of Second-order Stochastic Dominance criterion, using a 3 years rolling window and 6 market common trends, enhanced with quarterly cumulated "Price-targets" total news sentiment in black, the relative baseline reference strategy in red, and the market index in green.

The results (see figures 5.9, 5.10, 5.11, 5.12, and tables 5.2, 5.3, 5.4, 5.5, 5.6, 5.7) show that Second-order Stochastic Dominance and its scaled version seem able to better exploit the enhancement given by the news, while the optimization of the Sharpe Ratio produce a more modest enhancement. The reduced improvement for Sharpe Ratio could be due to the fact that volatility is the only risk measure considered to evaluate the portfolio performance. The baseline model residuals, and consequently the part fitted by the news, should be relatively small if compared to the total fitted volatility, and therefore the effect on the portfolio selection could result limited or irrelevant for several categories of news. The stochastic dominance strategies instead, being able to better exploit the full copula structure of the processed data, seems more adequate to exploit the advantages of the latent factors explained by the news.

The volatility and turnover of the strategies are almost unaffected by the news enhancement, therefore the discussion focuses on the final wealth despite it is not the only measure considered.

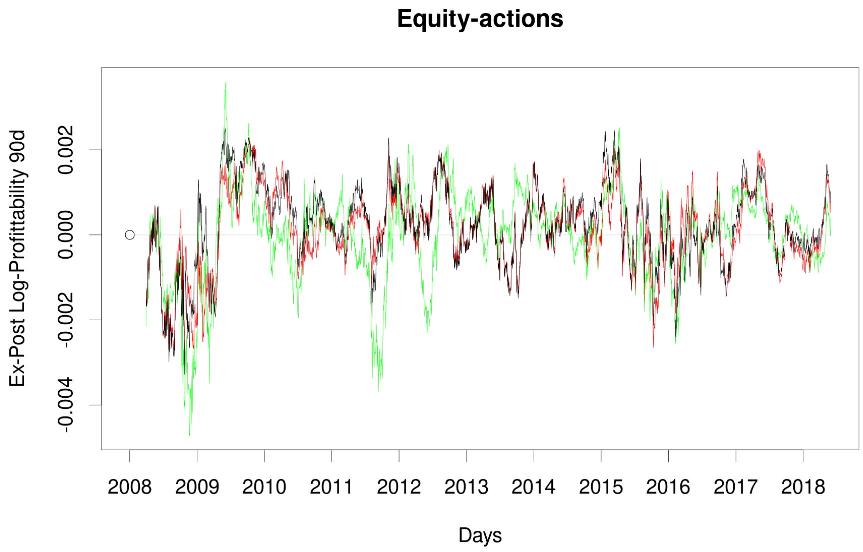

Figure 5.13.: Plot of the 90-days average profitability relative to the strategy reported in figure 5.11. Profitability of the news enhanced strategy (Weekly cumulated fresh "Techniacal Analysis") in black, of the baseline reference strategy (Scaled SSD with PCA) in red, and the EURO STOXX 50 market index in green.

A consistent portion of the considered category groups shows promising results, especially for the scaled version of the stochastic dominance. Nevertheless often the overperformance on the baseline strategy, are concentrated in small periods in which apparently the selected typology of news has a profitable influence.

The news enhancement seems to have almost no effect for the strategies based on the SSD criterion in the period between the second half of 2012 and the first half of 2014. A similar, less evident, period of strategies poorly affected by news is present also for scaled SSD in 2013. Some of the most profitable periods for scaled SSD are the recovery phase in 2009, between the second half of 2010 and the first half of 2011, and in the first half of 2018. The strategies based on Sharpe Ratio instead show limited periods of poor enhancement, mainly in the second half of 2008, in the first half of 2012, and in 2016.

In the tables 5.2, 5.3, 5.4, 5.5, 5.6, 5.7 the final wealth for each strategy is reported. Strategies reporting an increase of at least 0.15 are reported in bold in the tables and are considered good candidates. Weekly cumulated "assets" and "marketing" fresh news sentiment strategies, monthly cumulated "earnings" and "revenues" total news sentiment strategies, monthly cumulated "price-targets" fresh news sentiment strategies, and quarterly cumulated "dividends" total news sentiment strategies are good candidates both the stochastic dominance criteria, while the best performing strategies are weekly cumulated "equity-actions" fresh news sentiment (figure 5.11), quarterly cumulated "credit" fresh news sentiment, and weekly cumulated "stock-prices" total news sentiment.

Table 5.2.: This table shows the final ex-post log-wealth obtained enhancing with weekly cumulated total news sentiment from different category groups the 3 considered baseline strategies: Sharpe Ratio optimization, Second-order Stochastic Dominance, and the scaled version of SSD.

Category Group	**Sharpe Ratio**	**SSD**	**Scaled SSD**
acquisitions-mergers	-0.786	-0.858	0.443
analyst-ratings	-0.838	-1.116	0.335
assets	-0.786	-1.043	**0.542**
credit	-0.745	-0.858	**0.604**
credit-ratings	-0.736	-1.118	0.477
dividends	-0.783	-0.857	0.255
earnings	-0.736	-1.115	**0.638**
equity-actions	-0.803	-1.061	0.331
labor-issues	-0.767	-0.908	0.392
legal	-0.783	-0.985	0.277
marketing	-0.809	-0.87	0.456
partnerships	-0.771	-0.994	0.391
price-targets	-0.788	-0.9	**0.534**
products-services	-0.715	-0.923	0.499
revenues	-0.754	-0.982	0.477
stock-prices	-0.677	-0.987	**0.692**
technical-analysis	-0.828	-1.061	0.326

Table 5.3.: This table shows the final ex-post log-wealth obtained enhancing with weekly cumulated fresh news sentiment from different category groups the 3 considered baseline strategies: Sharpe Ratio optimization, Second-order Stochastic Dominance, and the scaled version of SSD.

Category Group	**Sharpe Ratio**	**SSD**	**Scaled SSD**
acquisitions-mergers	-0.804	-0.958	0.437
analyst-ratings	-0.831	-0.9	0.34
assets	-0.8	**-0.767**	**0.573**
credit	-0.784	-0.994	**0.536**
credit-ratings	-0.785	-1.041	**0.556**
dividends	-0.769	-1.024	0.45
earnings	-0.762	-0.868	0.468
equity-actions	-0.767	-0.904	**0.766**
labor-issues	-0.714	-0.854	0.505
legal	-0.736	-0.828	0.438
marketing	-0.739	**-0.802**	**0.552**
partnerships	-0.754	-0.988	**0.592**

price-targets	-0.8	-1.003	0.483
products-services	-0.741	-0.848	0.361
revenues	-0.743	-0.871	**0.537**
stock-prices	-0.708	-0.855	0.398
technical-analysis	-0.814	-0.856	-0.028

Table 5.4.: This table shows the final ex-post log-wealth obtained enhancing with monthly cumulated total news sentiment from different category groups the 3 considered baseline strategies: Sharpe Ratio optimization, Second-order Stochastic Dominance, and the scaled version of SSD.

Category Group	Sharpe Ratio	SSD	Scaled SSD
acquisitions-mergers	-0.748	-0.911	**0.645**
analyst-ratings	-0.742	-1.108	0.34
assets	-0.77	-1.093	0.329
credit	-0.79	-1.172	0.479
credit-ratings	-0.731	-0.888	0.448
dividends	-0.796	-0.903	0.509
earnings	-0.792	**-0.696**	**0.526**
equity-actions	-0.774	-1.005	0.362
labor-issues	-0.731	-0.832	0.473
legal	-0.774	-1.047	0.402
marketing	-0.776	-1.299	0.367
partnerships	-0.724	-0.936	0.354
price-targets	-0.722	**-0.795**	0.245
products-services	-0.787	-0.856	0.379
revenues	-0.807	**-0.794**	**0.585**
stock-prices	-0.761	-0.92	0.518
technical-analysis	-0.79	-1.177	0.41

Table 5.5.: This table shows the final ex-post log-wealth obtained enhancing with monthly cumulated fresh news sentiment from different category groups the 3 considered baseline strategies: Sharpe Ratio optimization, Second-order Stochastic Dominance, and the scaled version of SSD.

Category Group	Sharpe Ratio	SSD	Scaled SSD
acquisitions-mergers	-0.759	-0.866	0.346
analyst-ratings	-0.767	-0.934	0.341
assets	-0.795	-0.957	0.445
credit	-0.803	-0.929	0.512
credit-ratings	-0.763	-0.978	0.271
dividends	-0.797	-0.91	0.444
earnings	-0.811	-1.044	0.33

equity-actions	-0.799	-0.821	0.247
labor-issues	-0.781	-0.869	0.361
legal	-0.795	-0.883	**0.583**
marketing	-0.701	-1.056	0.346
partnerships	-0.706	-0.861	0.44
price-targets	-0.713	**-0.723**	**0.534**
products-services	-0.772	-1.048	0.069
revenues	-0.789	-0.852	0.478
stock-prices	-0.737	-1.021	0.486
technical-analysis	-0.782	-0.99	0.444

Table 5.6.: This table shows the final ex-post log-wealth obtained enhancing with quarterly cumulated total news sentiment from different category groups the 3 considered baseline strategies: Sharpe Ratio optimization, Second-order Stochastic Dominance, and the scaled version of SSD.

Category Group	Sharpe Ratio	SSD	Scaled SSD
acquisitions-mergers	-0.772	-0.948	0.391
analyst-ratings	-0.693	-1.075	0.384
assets	-0.726	-0.869	0.388
credit	-0.741	-0.917	0.487
credit-ratings	-0.731	**-0.786**	0.408
dividends	-0.788	**-0.803**	**0.588**
earnings	-0.768	-0.857	0.322
equity-actions	-0.781	-1.021	0.377
labor-issues	-0.762	-1.125	0.492
legal	-0.783	-1.064	0.474
marketing	-0.783	-0.884	**0.542**
partnerships	-0.719	-0.925	0.338
price-targets	-0.752	-0.901	**0.634**
products-services	-0.804	-0.825	**0.524**
revenues	-0.793	-0.897	0.422
stock-prices	-0.737	-1.138	0.194
technical-analysis	-0.775	-0.958	0.139

Table 5.7.: This table shows the final ex-post log-wealth obtained enhancing with quarterly cumulated fresh news sentiment from different category groups the 3 considered baseline strategies: Sharpe Ratio optimization, Second-order Stochastic Dominance, and the scaled version of SSD.

Category Group	Sharpe Ratio	SSD	Scaled SSD
acquisitions-mergers	-0.775	-1.035	0.37
analyst-ratings	-0.719	-0.936	0.297

assets	-0.7	-0.981	0.342
credit	-0.726	-0.863	**0.7**
credit-ratings	-0.786	-0.84	0.332
dividends	-0.784	-0.948	0.434
earnings	-0.765	-0.848	0.467
equity-actions	-0.781	-0.981	0.482
labor-issues	-0.765	-0.992	0.318
legal	-0.765	-0.953	**0.63**
marketing	-0.781	-0.843	0.412
partnerships	-0.734	-0.871	**0.566**
price-targets	-0.769	-1.018	0.479
products-services	-0.767	-1.023	0.301
revenues	-0.777	-0.901	0.476
stock-prices	-0.812	-0.907	0.409
technical-analysis	-0.792	-0.93	0.307

Table 5.8.: This table summarizes the results for scaled SSD reporting the selected category groups for each cumulation period and sentiment weighting considered: weekly total, weekly fresh, monthly total, monthly fresh, quarterly total, and quarterly fresh news sentiment.

Category Group	W.T.	W.F.	M.T.	M.F.	Q.T.	Q.F.
acquisitions-mergers			**+**			
analyst-ratings						
assets	+	+				
credit	+	+				**+**
credit-ratings		+				
dividends					+	
earnings	**+**		+			
equity-actions		**+**				
labor-issues						
legal				+		**+**
marketing		+			+	
partnerships		+				+
price-targets	+			+	**+**	
products-services				**-**	+	
revenues		+	+			
stock-prices	**+**				-	
technical-analysis		**-**			-	

A summary of the results for the news enhanced scaled SSD strategy is reported in table 5.8. In the table, a + is present if the news based strategy increases the final wealth of the baseline strategy of at least 0.15, and a bold + is present if the increase is of at least 0.25. Symmetrically a - and a bold **-** are present if the final wealth decrease is respectively at least 0.15 and 0.25. The results show that in general, the news enhancement has a positive effect reporting 24 plus and just 4 minuses. The best performing cumulation period seems to be weekly, having 12 plus and just 1 minus. The news freshness seems to be an important parameter for the analysis, having 12 plus and just 2 minuses. The weekly cumulated fresh sentiment is the indicator class showing the most promising performances. The category groups “credit” and “price-targets” both report the largest number of profitable strategies, equal to 3.

The same kind of analysis has been performed also on daily cumulated news showing profitable results in line with these reported in the previous pages. The results for daily cumulated fresh news sentiment reported the largest number of strategies improving the baseline, 9, confirming the goodness of the freshness weighting for short cumulation periods. The category groups improving the baseline for daily cumulation are “earnings”, “revenues”, and “technical-analysis” for total news sentiment and “acquisition-mergers”, “analysts-ratings”, “credit”, “dividends”, “earnings”, “labor-issues”, “price-targets”, “revenues”, and “stock-prices” for fresh news sentiment.

5.6. Further improvements

The proposed models are attractive and produce interesting performances, nevertheless, many improvements are known and could be applied to improve the performances of the models.

5.6.1. PCA refinement

One of the main tools, other than PCA, in the field of Exploratory Factor Analysis (EFA), is the Principal Axis Factorization (PAF). Replacing the PCA with PAF may improve the performance of the models because the PAF iteratively tries to improve the direction of the principal components returned by the PCA taking into account the percentage of variance explained by the principal components considered.

[OT15] propose and compares a wide set of different correlation measures that may be applied instead of the classical Pearson’s linear correlation. The authors define a φ-correlation measure as the Pearson’s linear correlation between a monotonic function of the original values, preserving some properties. A φ-correlation measure may be obtained mapping the quantile of the returns empirical distributions, considered on a long period, on these of a normal distribution.

To reduce noise in the original returns and better fit possible lead-lag effects, temporal smoothing may be applied to the original time series. Extreme returns outcomes may be considered as outliers or underweighted to improve the results, as these are considered to ill-condition the PCA and to introduce noise in the process.

5.6.2. Hyperparameters optimization

The considered model employs at least 5 different lookback windows: One to evaluate the market principal components, one in the market common trends regression, one to evaluate the news principal components, one in the news regression, and one to evaluate the portfolio optimization criteria. For the sake of simplicity, all these lookback windows are kept equal to each other. The possibility to inspect the optimal length of these windows may be of interest.

Moreover, also the optimal number of market common trends and of news factors to apply in the regression could be seen as a hyperparameter to further improve. The optimal number of news factors could also differ between category groups because different news categories could vehiculate a different amount of information.

Another possible hyperparameter that may be taken into account to improve the model performance is a positive or negative temporal lag on the effect of the news. The lag could be useful because news cumulation alone could not be enough to reflect the discrepancy between news publication and effects. The effect could anticipate the news due for example to information leakages, or the news could be discounted later, for example at earning announcements. The cumulation process could be enhanced by exploiting a Gaussian smoothing, on the entire cumulation window or just on the windows hedges, instead of an equally weighted vanilla cumulation. Information about the news effect temporal lag could be extracted by analyzing a lagged version of the naive beauty context developed in section 4.1, where the final wealth at each temporal lag may signal the presence of the news effect, and then the sequence of positive and negative lags where the effect is present may define the cumulation window. The cumulation window extracted in this way could also be dynamic, exploiting a rolling lookback window as previously done for the market common trends. Due to a possible ripple effect generated by the news on the stock returns the possibility to consider the residuals in absolute value for the regression and then to report the original sign should also be considered.

5.6.3. Mix of news factors

Once detected the category groups of interest for the model, the most correct way to mix them may be analyzed. Since the news has to be regressed on the baseline model residual as factors, mixing them up could not be easy. One possibility may be to use a weighted sum of different news category groups from the beginning before to compute the news principal components. Another possibility may be to select the optimal composition of factors to use in the regression. Due to the curse of dimensionality, as many category groups are present and because the optimal number of factors should be tested for each combination, to find the optimal solution could be computationally costly and time-consuming. The best way to search for this kind of solution could be an incremental process, where new factors are added one a time to the best solution found.

Another issue that could be taken into account is that category groups seems to affect return in particular periods, maybe in response to external macro-economic events or market conditions, and that probably some market sector may be more affected than others. For instance, negative outcomes of the principal factors for "'credit-ratings"' could affect the returns of the banking sector, as a general rating decrease of other companies for in place credits could affect perspectival banking earning estimates.

5.6.4. Regression regularization

The regression of the returns on the market common trends, and of the residuals on the news principal components is subject to the random effects of noise and spurious factors regression (see figure 5.1). Many authors have focuses their studies on the possibility to reduce these effects through regularization techniques. Some of the more common regularization techniques include Lasso and Ridge regressions, respectively L_1 and L_2 regression weights normalizations, ElasticNet regularization that linearly combine the two previous methods, and Stepwise regression, a model selection technique based on the Akaike Information Criteria (AIC). Some preliminary studies have been conducted in this work using the Stepwise regression but the computational complexity of the model selection makes the regression highly time-consuming.

5.6.5. Scenario generation

Preliminary results show that the considered source of news sentiment and its principal components do not seems to be a profitable state driver for scenario generation. This result could be due to a miss-specification of the risk scenario, given by the removal of one side of the supposed swinging market structure (i.e. bear and bull market structure) from the distributions, while SSD and its variants may need to exploit the whole sample. Anyway, the capabilities of more specific and possibly sector-specific news indicators have still to be analyzed.

5.6.6. Transaction costs and turnover optimization

The portfolio strategy obtained seems to clearly outperform, in terms of returns, the market index in the considered period. Anyway, no information about transactional costs, liquidity problems, or market moving against the trades is available in the database or is considered in the results at a first stage. Some theoretical studies have shown that imposing liquidity constraints, limiting the availability of each stock by a fraction of the recent average stock trading volume, the performance of portfolio strategy, especially news-based portfolio strategy ([HLGC+19]), could suffer a sensible decrease in performance when the invested capital exceeds 10 million dollars and approaches 1 billion dollars.

5.6.7. Different turnover models

A more reliable portfolio strategy need to lower turnover in search of minimizing transaction costs. Different enhancements could be considered, stabilizing the portfolio weights through time, limiting the turnover ratio, or just choosing to not rebalance up to a predetermined calendar date.

Preliminary results show that with a rebalancing period of 5 days, turnover and therefore, at least a part of, transaction costs are reduced, but the over-performance given by the market common trends seems to be reduced and less effective in respect to the original signal. The over-performance of market common trends on returns different from daily returns instead has still to be investigated. The analysis of returns different from daily may be of interest because these could result appropriate for rebalancing periods longer than daily.

Initially, the portfolio weights solving the optimization problem and used to allocate a portfolio were kept fix for a predefined period of time. The considered periods were 7 and 30 days. After that period, portfolio rebalancing was allowed, and the rolling window moved ahead. The choice to rebalance the portfolio only every 7 or 30 days were taken to reduce the transactional costs that in an active portfolio strategy could turn a profitable strategy into consistent losses.

Despite the possible increase in transaction costs due to a higher turnover, in this work we opted for a daily rebalance, allowing a more in-depth study of the portfolio turnover performance, and leading to a more conspicuous sample size to better evaluate portfolio performance. Thus each day the rolling window is updated and the portfolio rebalance is allowed, but only for stocks making a price on that day. Stocks without a price are considered in a non-tradable state and the positions are kept fixed.

5.6.8. News enhanced forecast model

A different model has also been preliminarily evaluated but kept aside in this work, and hereafter briefly reported, especially due to issues in the exact and complete portfolio components covariance estimation.

To enhance the baseline portfolio strategy, a set of “views” was generated from the cumulated news indicators previously described. The model, also if very different from Black-Litterman, shares with it some similarities in terms of processing views. A view is a statement on the market, the Black-Litterman model considers views on expected returns. In Black-Litterman, views are expressed as normal distributions where mean and variance respectively quantify the views and the uncertainty and the posterior distribution is computed using the Bayes' formula.

The views generated using the news cumulated indicators instead do not quantify the expected returns of a stock but the excess return and volatility against the baseline strategy. The forecast of the return distribution was extracted from the residuals of the market common trends regression. Such kinds of views are more easily extracted from the data and applied to the strategy, due to the nature of the model. The equivalent of the posterior distribution in the Black-Litterman can thus be easily computed adding the forecast of the excess return and variance to the expected return and variance of the generated scenarios and then used to compute the Sharpe Ratio as done in the baseline strategy. The excess variance can be easily added to the generated scenario variance, but the calculation of its covariance term results computationally hard, and two possibilities have been hypothesized: the computation of the full variance-covariance matrix of the forecast, or the re-estimation of the forecast for the current portfolio composition at each iteration step in the portfolio weights optimization. The views on volatility may also be applied as an innovation term on the actual volatility level in GARCH like models.

Nevertheless, the application of views in a stochastic dominance framework remains a problem and needs further research. The reason is that, as stochastic dominance is able to exploit the latent information hidden in the copula of the multivariate returns distribution, the way in which a view has to be applied to each one of the samples in the return distribution remain undefined. We named the transformation of the inputs, the prior returns distribution and the view, into the output, the posterior return distribution, the "transfer function" in line with the transfer function used in statistical time series analysis, a mathematical relationship between the numerical input to a dynamic system and the resulting output. The most simple transfer function for views on excess returns is a distribution shift proportional to the strength of the views, obtained by adding a fraction of the view strength value to each sample composing the distribution. In the case of views on volatility, a simple transfer function could be a stretch of the distribution proportional to the strength of the view, obtained by multiplying each sample in the distribution by a value proportional to the strength of the view, while keeping the mean fixed. Anyway, there is no guarantee, especially for views on volatility, that these transfer functions are suitable for the problem and thus generate consistent results. The view effects on the copula structure may be unpredictable, because, for example, the view may try to statistically predict the presence of events in the near future that have no correspondence in the historical time series, such as in extreme sentiment situations or when the sentiment is out of the historical range, and therefore a suitable and exact transfer function may even not exist. Anyway, also when a suitable transfer function is not identified, these views could be exploited in a portfolio optimization framework for tactical asset allocation tilting, overweighting or underweighting stocks according to views, or also for stocks pre-filtering before the portfolio optimization.

6. Conclusion

6.1. Pre-opening news portfolio selection

The fourth chapter of this work analyzed the effects on the Euro Stoxx 50 market index components of pre-opening news indicators. The pre-opening indicators are calculated as 4 hours pre-opening cumulated firm-specific news sentiment or volume. The indicators have shown interesting performances when applied in portfolio strategies. The news event taxonomy resulted to play a central role for the strategies. The news freshness instead seems to do not be highly relevant for the pre-opening intraday strategies, except for few specific situations. Three main patterns emerge from the data. The first pattern is related to well know proverbial phrases (i.e. “No news is good news” and “Sell on the news”), covers the largest group of category groups and is based on selling stocks with a large volume of news. Generally, this pattern shows reversal overnight. The second pattern allocates the portfolio according to the total cumulated news sentiment, buying stocks with positive sentiment and selling the ones with negative sentiment. This pattern shows a weak reversal effect overnight for most of the category groups, except for “Technical analysis” that show strong reversal overnight. The third pattern allocates the portfolio according to the inverse of total cumulated news sentiment, selling stocks with positive sentiment and buying the ones with negative sentiment. This pattern shows a moderate reversal effect overnight for most of the category groups.

The analysis shows clear profitability for many strategies related to the data. The strategies characterized by overnight continuation effect, or at least not characterized by a strong reversal effect, may be good candidates for longer horizon holding periods (i.e. positive “analyst-rating”, total “earnings”, total “price-targets”, and total “products-services” sentiment).

6.2. Portfolio selection with news

The fifth chapter of this work analyzed portfolio strategies enhanced by the news and based on three different optimization criteria. Once again the results are evaluated on the components of the Euro Stoxx 50 market index, and the index is taken as reference. The strategies rely on a baseline model in which returns are regressed on market common trends extracted applying the principal component analysis. The news enhanced strategies are then obtained regressing the residuals of the baseline model on the news principal components extracted from each news category group separately.

The results for the baseline strategies shows that the Scaled Second-order Stochastic Dominance is the best performing optimization criteria on the considered period, including the 2008-2009 financial crisis. The strategy obtained with the Scaled SSD criterion outperforms the market index in terms of final wealth and volatility. The Sharpe Ratio and Second-order Stochastic Dominance strategies instead suffer severe losses during the 2008-2009 financial crisis and do not revert to profitability until 2011. Both the strategies underperform the market index in terms of final wealth and volatility. The Sharpe Ratio strategy is characterized by a period of high profitability in 2015. The SSD instead is, counterintuitively, the strategy achieving the largest drawdown. The Sharpe Ratio strategy is the one with the lower turnover, while the stochastic dominance strategies achieve more than two and three times its turnover level, with the Scaled SSD strategy reaching the higher turnover level.

The news enhancement effect result to be limited when the Sharpe Ratio criterion is applied, while better results are achieved applying the stochastic dominance criteria. Volatility and turnover resulted to be almost unaffected by the news enhancement. The results have then been mainly evaluated on the final wealth achieved by the strategies. The news enhancement resulted to be positive in general, with many strategies outperforming the baseline model final wealth. The analysis has shown that profitability tends to concentrate in small periods making the results of the strategy more chaotic and difficult to validate. Some of the news category groups shown promising results for both the stochastic dominance criteria using the same weighting and cumulation period. The news freshness resulted to be an important factor to consider, producing profitable strategies especially over short cumulation periods. Fresh news daily and weekly cumulated indicators resulted to be the best candidates for profitable strategies based on category groups, also if profitable strategies are present also for longer cumulation periods, as quarterly cumulation.

Acknowledgments

The authors acknowledge Optirisk Systems Limited, UK's support. The Company Specific News Meta Data and the Market Data have been supplied by OptiRisk Systems Limited. These contributions are gratefully acknowledged.

The authors thank for Italian grants ex-MURST 60% 2020 and acknowledge the support of the Czech Science Foundation (GACR) under Project 20-1674S.

A. Firms, prices, investors, and news

This chapter describes real events reported by news media that can affect market prices and thus portfolio management. The information hereafter reported is mainly extracted from the database used that is better described in chapter 3. The description of the real-world economy, performance evaluation, and partially of equity market takes as reference the structure of the RavenPack taxonomy as found in the database. Some other contributions have been added by the author, and are relative to his own analysis of the markets or of its contingency in the last years.

A.1. The real-world economy

A.1.1. The value generation process

A firm is an entity having the aim to generate value for its stakeholders, such as customers, employees, and investors. To create value, it employs people and assets in its processes.

Employed persons belong to a labor market that can have different characteristics in different countries and geopolitical areas. Employees, executives, but also board members, are people who have to be searched for, who have a salary, who could die, resign, retire, or be fired, and who could also be involved in scandals. The labor market to which they belong is (normally) populated also by other players such as regulators, the government, and labor unions. The regulator could change the labor policies, implying different salary or firing requirements, while labor unions could require a union pact or call for a strike.

To produce goods and services by consuming production factors, in addition to employees, different kinds of assets are needed. Each one of these assets has an implicit value that can vary over time, and if needed sometimes these can also be sold. The assets include, but are not limited to:

- Facilities that could be opened, closed, upgraded, or relocated and that have maintenance costs.
- Commodities that could be bought, sold, or offered.
- Patents that could be awarded, revoked, or could also expire.
- Entire units or companies.

The outputs of the process are products and services. These are characterized by a life cycle that could also be related to its marginal gain. Each different product or service can be in a different stage of its life cycle, such as development, release, enhancement, or recall. A product could also be discontinued, abandoned, or resumed. Promotions could be applied, side effects could be discovered or reviews with specific sentiment could be done. Each product and service is also, almost always, characterized by a price that could be raised or cut. Some products are subject to regulatory approval. An application to the approval is required, and the approval could be granted, conditional, or denied. Others are subject to patent enrollment or expensive and complex verification processes, such as clinical trials in the pharmaceutical industry. Firms seek to break into markets where they can gain or lose shares. For each product, a market has a demand that varies over time that firms try to satisfy with an optimal level of time-varying supply.

To finance its overall operations and growth a firm relies on a capital structure that is composed of equity and debt. The equity capital is conferred to the firm with the initial capital and can be increased by retaining earnings or by a capital increase, a complex procedure that has to be approved by the board of directors and in which new equity shares are issued for capital. The debt capital is expected to vary over time and could sometimes need to be restructured, a procedure subject to approval. It is constituted of loans, normally issued by banks, and capital coming from the issue of bonds. Bonds once issued by firms are collocated on the primary market and can then be traded on the secondary market. The trading price of bonds fluctuates over time but tends to converge to the established value approaching maturity dates unless bankruptcy occurs. The possibility to collect equity capital and collocate bonds depends also on the market risk premium, as investors could choose to invest in different opportunities if these seem to better reward the same level of risk. The possibility to borrow money depends on other investment opportunities too, but also on the liquidity cost set by the regulators, the central bank, through the active and passive interest rates.

Another important factor that has an impact on firms is the judicial system. These in the various degrees of judgment are subject to verdicts and sanctions.

The value generation process could be also affected by external events such as corruption, crime, civil unrest, industrial accidents, security issues, war conflicts, or, as it is happening these days, pandemics and lockdowns.

A.1.2. The firms' interconnection

Firms are not isolated and standalone entities, these populate a highly interconnected system. The outermost entity in which a firm is inserted is a geopolitical area. A geopolitical area is a concept with a non-unique definition, it can coincide with a country, a group of countries, or more generally a portion of a continent. Different geopolitical areas and commonly different countries are characterized by differences in the regulation of the labor market, in the judicial system, and in the type of government, but also by different currencies, central banks policies, alliance systems, and expenditure repayment dynamics. A firm is normally initially based in one country and then possibly spreads to others, few are born international. When a firm trades across borders, the transactions are classified as import and export and these transactions affect the balance of payments.

Each firm participating in a market sector is subject to Porter's five forces. Three forces represent "horizontal" competition and are substitute products, competitors, and new entrants, while two forces represent "vertical" competition and are suppliers and customers. The supply chain provides the inputs to the firm and plays an important role in the value generation process. The customers, who can be divided into different market segments and differentiated between consumer and business, represent the demand for the firm supply, which can be time-varying and affected by external factors, such as propensity to consume, inflation, seasonality, and economic cycle, but also by factors driven by the firm, such as marketing. Substitute products, competitors, and new entrants populate the market sector that is characterized by entry and exit barriers. The crowding of the sector, the size of the firms, the market shares controlled by each player, and the entry and exit barriers are some of the elements defining the type of competition in the market that can vary from monopoly to perfect competition. In the case of perfect competition due to low barriers, firms should focus on scale economies or diversification (market niches) to maintain a good marginal gain and avoid a price war. The barriers can be established by costs, (i.e. of production, or of entry into the market), but, in regulated markets, also by the regulator or the government that, when needed, could modify them over time (i.e. with deregulation).

Within the market to improve their performances firms can establish "horizontal" or "vertical" partnerships, which could be opposed, for example by the antitrust regulation authority.

A firm could also participate in acquisition, merger, private placement, or public offering. In these operations firms or units are valued with variable bidding, and, as for partnerships, also these operations could be opposed and are subject to regulatory approval.

A.2. Performance evaluation

A.2.1. Fundamental indicators

Periodically firms engage in investor relations and publicly release fundamental indicators within their financial statements. Normally listed companies release financial statements quarterly.

The main indicators reporting generated value are inventory, interest income, revenues, earnings, and dividends. Inventory reports the quantity of goods acquired but not consumed by the process. Interest income represents the amount paid to an entity for lending its money or letting another entity use its funds. Revenues, which can be subdivided into operating margin, revenue, revenue volume, and same-store sales, report the incomes generated before subtracting expenses. Earnings, which can be subdivided into earnings, earnings per share, EBIT, EBITA, EBITDA, operating earnings, and pre-tax earnings, report the profit a company has earned. It is calculated by subtracting expenses, interest, and taxes from revenue. Earnings announcement for listed companies normally comes at scheduled calendar dates. Dividends are the distribution of a part of the earnings to some class of its shareholders.

All these fundamental indicators normally represent a lagged indicator because these are publicly released only at the end of the considered period, and often with some additional lag.

The indicators of two different firms can not always be easily compared because each different market sector is characterized by a different optimal level for each one of the indicators.

To achieve a wider view of the economic situation different institutions also produce aggregated indicators about market sectors, countries, and geopolitical areas. Some of the most relevant aggregated indicators are gross domestic product, inflation, consumer spending, house pricing.

A.2.2. Perspectival views and ratings

The evaluation of a firm is not only related to its current capacity to generate value. Instead, the value generation capacity has to be considered under a perspectival view. To understand the latent value the possible evolution and the growth capacity of a market, a sector, and of the firm itself have to be taken into account.

Financial analysts are specialist that study firms and its interconnection to better understand its latent value and its perspectival possible evolution. Analyst's ratings are normally expressed using a granular score suggesting the portfolio management action to take in respect to a firm. In addition to ratings, analysts, also report estimations on price targets and on the expectation of future values for the fundamental indicators, such as earnings, revenues, and dividends. Despite the highly regulated sector of analyst ratings, they are people working for financial firms or investment banks and sometimes their studies can suffer from agency problems.

The analyst's estimation techniques nowadays are so evolved that in some sectors satellite imagery and aerial photography are used to make automated estimates (see [DA20]). For instance studies on the analysis of car parking occupancy exist for retailer's revenues estimates. Other studies have shown that nighttime lights are good nowcasters for GDP growth, especially in emerging markets, and that tracking industrial activity proxied by the earth's surface reflectance can help in nowcasting PMI manufacturing indexes. Other estimation techniques include, but are not limited to, GPS (global positioning system) location, mobile phone location, taxi ride tracking, corporate jet location, readership of financial sources, search engine data, consumer transaction data, and consumer receipts data. For instance, GPS and AIS data can be used to monitor vessel traffic to understand flows of crude oil, while mobile phone location and consumer transaction data can be used to understand retail sales activity.

Similarly to analyst rating, a credit rating is a granular score that assesses the creditworthiness of a borrower, the capacity of a firm to repay its debts, and is normally issued by a rating agency. Rating agencies are specialized companies that, employing analysts, periodically publish assessments on the ability of firms and governments to generate value and repay debts. Regulators in different countries can set a minimum rating score, defined investment grade, below which some kind of institutional investors could not be allowed to hold the debt security. Among the most eminent rating agencies are S&P Global Ratings, Moody's Corporation, and Fitch Ratings.

Economic shocks can be produced by unpredicted events. One reason for these shocks could be that these events break plans and suddenly change perspectival views. Firms, which have to carry out mid and long-term business plans, incur costs to change their plans in progress and lose their planning advantage, due to having relied on perspectival visions that have not been realized correctly when the uncertainty was removed.

A.3. Equity market

A.3.1. Ownership structure

When a company is established, an initial capital is conferred to it. Persons and institutions conferring the capital are the company owners. In joint-stock companies, a precise number of shares are matched to the capital conferred and these are proportionally divided between the owners. The company shares may be bought or sold by shareholders, these transitions are said to be "over the counter".

To raise capital a company can place part of its shares on a regulated market. The placement on the regulated market is called Initial Public Offering, the shares are quoted and sold on the primary market to shareholders. Any further trade of the already placed shares take place on the secondary market and do not affect the capital of the company. The gain and loss on the secondary market only affect shareholders portfolios.

When a company is listed on a regulated market it has to comply with stricter regulations, implying more transparency, more solidity, and periodical information disclosure to the public. Some regulated markets also require the nomination of specific advisors to fulfil particular functions. The listing ensures more liquidity to the stock for trading purposes, a consensus on the stock price is publically available (if enough liquidity is present), and offers can continuously arrive on the market. The buy and sell offers arrive on the market platform and are listed on the limit order book. When the price of a buy order matches the price of a sell one, a transaction is made, the price is reported as the consensus on prices, and the transacted quantity is added to the transaction volumes. Due to the order flow prices can freely move up or down. There is an entire field studying the prices movements denominated technical analysis. The main information extracted regards the strength of price movements, price levels for resistances and supports, bullish or bearish views, and overbought or oversold situations.

The Relative Strenght Index (RSI), and technical price levels are considered by technical analysts to determine overbought or oversold situations, and bearish resistance or bullish support price levels. Technical analysts also produce views regarding bearish or bullish market trends and overbought or oversold situations.

If too many buy or sell orders are placed concurrently order imbalances can take place, in some situations the trading of a stock can be halted due to imbalances.

When needed the capital can be raised with a capital increase, a secondary public offer taking place on the primary market, placing newly issued shares (dilutive) or placing all or a portion of the holdings of major stockholders (non-dilutive). The share can be subject to stock splits and reverse splits, affecting the total number of shares, and in some cases the liquidity, but normally leaving unaffected the total value of the shares.

Most solid, capitalized, and liquid stocks can also be listed in or delisted from specific market indexes.

A.3.2. Market players and the news

The order flow is sent to the market by investors, the market players. As highlighted by the Adaptive Market Hypothesis different kinds of investors exist and populate the market. Investors can be subdivided into retail and institutional. Retail investors are non-professional individuals, while the term institutional refers to different entities, such as pension funds, banks, market makers, hedge funds, insurance companies, and mutual funds.

Buy and sell orders received by the market platform reflects the investor's money flow. Investors allocate their money according to their beliefs. The investor's beliefs, except for noise traders, are supposed to be fed by the information flow, which should be based on real-world events. The investors get in touch with real events through the news, which is vehiculated by news media. The news, which is not always trusty, may also contain a certain level of uncertainty and may be transmitted with a consistent lag. **Each one of the aspects regarding firms previously described in this chapter is a real-world event that can be conveyed through the news.** Some of the reported events have a small impact on the stock prices, others have a big one.

One of the events, reported by the news, able to modify or at least affect investor's beliefs, and then the choice, are changes. Changes in values can affect:

- Fundamental indicators (such as earnings, revenues, and dividends).
- Assets.
- Import and export.
- Debt.
- Interests.
- Expenses.
- Ratings (such as analyst ratings, credit ratings, or price targets).
- Ownership.
- Aggregated indicators (such as gross domestic product, inflation, house prices, consumption propensity, and salaries).
- Products demand and supply.

To anticipate the effects of these changes analysts also produce estimates of these possible changes, as expectations on values. News pieces can also report information about the miss-estimation of these expectations. Miss-estimation could represent small shocks because investors could have taken a position on the market according to analysts' estimation, regardless of the estimation uncertainty, and then the difference between the estimate and the outcome could represent a break in the financial planning of the investor, who could have to modify the assets allocation, to reconcile the portfolio with the outcome. Outcomes above (below) the estimated level for an extended period of time could also represent an overperforming (underperforming) indicator of the company, due to latent factors overlooked in the analyst evaluation process. Analysts commonly produce expectations about:

- Fundamental indicators (such as dividends, earnings, and revenues).
- Aggregated indicators (such as domestic product).
- Products demand and supply.

News pieces regarding analysts ratings and credit ratings should be of interest to investors, especially when a rating changes, but also when it reports historical views because these pieces of information are generated by qualified professionals, that spent a lot of time trying to understand the real market situation. In some situations, these ratings could also force some kind of investors to exit from their existing positions, such as when a non-investment grade rating is assigned to a stock hold by particular kinds of institutional investors.

News can also vehiculate relevant information about events happening in the market, such as acquisition, merger, initial public offering, equity restructuring, and product development. Some investors could try to take a position on the market as soon as rumours or interests are published, also if uncertainty about the event final outcome, often associated with volatility, has not been resolved and thus volatility could still be rising. These market positions can also be modified when more information becomes available, as during price settlement, oppositions, and regulatory concerns. Conversely, other investors take positions when the operation is completed or failed and the uncertainty raised by the operation is resolved, and thus the volatility should be reduced.

News pieces regarding insider trading and buybacks are of interest because they could reflect latent beliefs of investors, that respectively at the same time take part in the management and ownership processes. Buybacks are normally initiated when the company stock price is considered to be low valued by the ownership and with a good chance to have positive perspectives, so it could be reputed positive also when regulatory concerns with negative sentiment are present. Insider trading instead could represent a positive, or negative, view given by the management. If the management, or the ownership, sell (buy) a considerable amount of his shares when the price of the stock is relatively high (low) this represent a strong (weak) negative (positive) view on the perspective possibilities to generate value. The positive signal could be considered weaker because the management could receive incentives to increase his/her shareholdings quota (for example to confirm rumours about a recovery from the lows), while a negative view could be the result of the stock quoting an un-realistic positive view on the company.

Investor beliefs can also be influenced by news regarding stock picks and marketing. Marketing is a relevant source of news because it is a field of research studying how to affect beliefs for a specific purpose. Marketing is not only a way to place industrial or consumer products, but also the financial industry can try to place its products. Marketing could regard products and services, in an attempt to increase demand, but also stock picks. Financial promoters are sellers with agency problems who do not always provide fully symmetric information. Financial products that are sponsored through a marketing process have to pay a cost on the risk premium trade-off. The marketing costs arisen from the sponsorship process has different origins. The costs have to be sustained by the targeted buyer investors, also for the unsuccessful part of the sponsorship. The costs represent a disadvantage also on short-selling products that are normally not sponsored by financial intermediaries because it could raise regulatory concerns. The costs are given by the sharing of the rewards on the risk premium with the marketing agency customer because the allocation is not optimized to maximize the investor's portfolio but the one in which agency-client profit is maximized. The marketing process consumes resources, initially to explain to customers what is a product and what is it for, but then it tries to push customers to buy products not optimal for the customer but for the marginal profit optimization of the marketing itself (For example starting a marketing competition, a “war”, between firms placing products on the same market). The marketing process could, in some extreme circumstances, also try to corrupt the news flow, in the attempt to drive the investors to the appropriate choices for its purposes, as by buying publication space on news media and inserting latent marketing messages in apparently unrelated information-carrying news pieces. While marketing often claims to self-regulate, in many cases, it does not seem to be happening.

According to behavioural finance investors, especially non-professional retail investors could be affected by biases, such as anchoring or availability, and marketing could easily exploit these inefficiencies. The availability bias could be exploited because the stocks proposed by a financial intermediary, a financial promoter, or by stock picking news are already present in the mental neighbourhood of the investor and thus could have an increased probability to be bought. The anchoring bias could then be exploited because, as the stock has already been bought, the investor could follow believing that his investment is a good investment, it can rely on it, do not have to be changed, and maybe that new investments can be made in the same stocks.

More sophisticated investors could modify their beliefs, and choices, also taking into account news regarding firms sustainability, environmental pollution and corporate responsibility. Also at a regulatory level, the Environmental, Social and corporate Governance (ESG) investments are taking place as a social need, that should be pursued by investors, especially institutional investors.

B. Additional exhibits

B.1. Statistics for pre-opening news indicators

Table B.1.: This table shows the statistics of 4 hours pre-opening cumulated news sentiment.

Category Group	Mean	Std.Dev.	Skewness	Kurtosis
acquisitions-mergers	0.024378	0.4176	32.0942	1440.3032
analyst-ratings	0.003030	0.1641	3.6068	246.1522
assets	0.000573	0.0488	1.0872	1734.7241
balance-of-payments	-0.000003	0.0009	-17.4209	3979.4812
bankruptcy	-0.000024	0.0016	-69.3173	4827.6527
civil-unrest	-0.000008	0.0006	-69.9500	4894.0002
corporate-responsibility	0.000032	0.0018	64.5424	4269.8844
credit	0.000397	0.0157	45.6135	2968.6672
credit-ratings	-0.000238	0.0216	-7.6580	3373.8591
crime	-0.000008	0.0005	-68.7061	4759.6347
dividends	0.000418	0.0380	21.6721	2667.5734
domestic-product	0.000000	0.0000	-	-
earnings	0.017119	0.4140	12.5562	533.6235
equity-actions	0.003592	0.1179	13.9475	1171.5594
exploration	0.000154	0.0052	35.5009	2016.0233
government	0.000000	0.0000	-	-
health	-0.000005	0.0004	-69.9500	4894.0002
housing	0.000000	0.0000	-	-
indexes	0.000023	0.0021	42.1790	4673.5891
industrial-accidents	-0.000077	0.0040	-39.9545	3323.2939
insider-trading	0.000009	0.0044	25.0449	4118.2958
investor-relations	0.000000	0.0000	-	-
labor-issues	0.000136	0.1437	-0.0612	1453.2130
legal	-0.002010	0.0669	-25.6373	2247.9485
marketing	0.000108	0.0049	56.9829	3555.7225
order-imbalances	0.000000	0.0000	-	-
partnerships	0.005537	0.1017	37.1288	1920.7551
price-targets	0.000857	0.0602	1.9525	455.0348
products-services	0.015904	0.1848	19.5089	896.9158

regulatory	-0.000546	0.0214	-47.3271	3647.3179
revenues	0.008543	0.1519	19.3908	1027.0574
security	-0.000032	0.0019	-63.0060	4092.6691
stock-picks	0.000036	0.0052	57.5207	4566.4493
stock-prices	0.000446	0.0971	0.1164	548.5700
taxes	0.000002	0.0001	69.9500	4894.0002
technical-analysis	0.000895	0.0451	1.5260	250.8991
transportation	-0.000055	0.0018	-37.8446	1516.5017
war-conflict	-0.000000	0.0019	-17.1091	4252.6606

Table B.2.: This table shows the statistics for 4 hours pre-opening cumulated fresh news sentiment.

Category Group	**Mean**	**Std.Dev.**	**Skewness**	**Kurtosis**
acquisitions-mergers	0.004340	0.0662	21.0718	724.2537
analyst-ratings	0.001997	0.1064	3.2257	173.1570
assets	0.000179	0.0146	3.5031	1520.6148
balance-of-payments	-0.000002	0.0005	-17.7066	3979.9786
bankruptcy	-0.000006	0.0004	-66.6266	4498.0953
civil-unrest	-0.000003	0.0002	-69.9500	4894.0002
corporate-responsibility	0.000008	0.0005	66.7591	4513.6336
credit	0.000115	0.0068	43.3021	2990.1930
credit-ratings	-0.000005	0.0076	1.5290	3462.8328
crime	-0.000004	0.0003	-69.9500	4893.9977
dividends	0.000279	0.0167	25.1564	2630.5849
domestic-product	0.000000	0.0000	-	-
earnings	0.009860	0.1804	15.2699	618.7394
equity-actions	0.000757	0.0317	10.0130	798.5511
exploration	0.000036	0.0012	39.9764	2042.6162
government	0.000000	0.0000	-	-
health	-0.000000	0.0000	-69.9500	4894.0002
housing	0.000000	0.0000	-	-
indexes	0.000008	0.0010	38.2197	4664.7586
industrial-accidents	-0.000010	0.0007	-44.0370	3733.3497
insider-trading	0.000011	0.0019	23.5309	4454.7861
investor-relations	0.000000	0.0000	-	-
labor-issues	0.000326	0.0273	3.7952	1008.2930
legal	-0.000238	0.0128	-24.2124	2310.1215
marketing	0.000025	0.0012	59.7216	3743.2372
order-imbalances	0.000000	0.0000	-	-
partnerships	0.000790	0.0191	30.0299	1542.2740
price-targets	0.000655	0.0516	1.0811	414.9892
products-services	0.002907	0.0424	16.6652	616.5489
regulatory	-0.000067	0.0040	-54.1162	4042.3285

revenues	0.003126	0.0658	16.3677	982.6761
security	-0.000010	0.0007	-69.9499	4893.9938
stock-picks	0.000037	0.0030	59.6733	4819.8325
stock-prices	0.000176	0.0288	5.5465	1719.9000
taxes	0.000000	0.0000	69.9500	4894.0002
technical-analysis	-0.000009	0.0103	5.6886	2114.4729
transportation	-0.000006	0.0002	-46.7114	2366.7691
war-conflict	-0.000006	0.0008	-17.4881	4893.9291

Table B.3.: This table shows the statistics for 4 hours pre-opening cumulated news volumes.

Category Group	**Mean**	**Std.Dev.**	**Skewness**	**Kurtosis**
acquisitions-mergers	0.049506	0.7548	31.5402	1366.9000
analyst-ratings	0.026410	0.2219	12.2561	218.2424
assets	0.008990	0.1597	32.2410	1337.8685
balance-of-payments	0.000102	0.0047	61.5484	3974.1487
bankruptcy	0.000033	0.0021	69.1025	4803.6115
civil-unrest	0.000039	0.0024	65.6853	4395.3615
corporate-responsibility	0.000120	0.0069	64.4991	4264.1721
credit	0.002831	0.0716	40.2369	1956.5691
credit-ratings	0.001429	0.0460	49.2450	2756.0697
crime	0.000030	0.0019	68.3463	4717.6650
dividends	0.006274	0.1389	34.8126	1521.8739
domestic-product	0.000006	0.0004	69.9500	4894.0002
earnings	0.101239	1.0966	18.6214	506.3930
equity-actions	0.023000	0.3150	26.6645	978.0024
exploration	0.000351	0.0086	40.4779	1991.8313
government	0.000000	0.0000	-	-
health	0.000009	0.0006	69.9500	4894.0002
housing	0.000000	0.0000	-	-
indexes	0.000057	0.0037	67.9801	4671.3431
industrial-accidents	0.000156	0.0079	55.2782	3299.1391
insider-trading	0.000150	0.0084	62.9176	4116.6982
investor-relations	0.003731	0.0893	43.7965	2287.7227
labor-issues	0.018926	0.3022	34.0071	1522.2158
legal	0.006478	0.1350	43.3905	2291.9696
marketing	0.001936	0.0425	54.6248	3353.5095
order-imbalances	0.000000	0.0000	-	-
partnerships	0.011905	0.2082	39.0547	1912.6045
price-targets	0.019665	0.1548	12.8265	238.7017
products-services	0.051202	0.4173	24.4353	876.1859
regulatory	0.001321	0.0473	57.2797	3545.7312
revenues	0.043565	0.4663	22.3771	744.9006

security	0.000057	0.0035	62.8905	4077.5003
stock-picks	0.000153	0.0097	67.0902	4556.1418
stock-prices	0.029715	0.3153	19.3889	537.3070
taxes	0.000003	0.0002	69.9500	4894.0002
technical-analysis	0.018482	0.1324	10.7324	160.0860
transportation	0.000129	0.0043	39.0709	1621.4092
war-conflict	0.000063	0.0034	64.4455	4263.9520

Table B.4.: This table shows the statistics for 4 hours pre-opening cumulated fresh news volumes.

Category Group	**Mean**	**Std.Dev.**	**Skewness**	**Kurtosis**
acquisitions-mergers	0.009194	0.1245	21.8490	681.6546
analyst-ratings	0.017053	0.1451	10.6971	147.9972
assets	0.002075	0.0423	28.2308	953.8633
balance-of-payments	0.000063	0.0029	61.6274	3995.9260
bankruptcy	0.000009	0.0005	66.5329	4485.7562
civil-unrest	0.000018	0.0010	63.1157	4077.5062
corporate-responsibility	0.000030	0.0019	66.7503	4512.7651
credit	0.000950	0.0276	43.2188	2221.5113
credit-ratings	0.000476	0.0177	51.8842	2974.0358
crime	0.000015	0.0011	69.9500	4893.9967
dividends	0.003110	0.0666	31.0018	1213.4679
domestic-product	0.000000	0.0000	69.9500	4894.0002
earnings	0.043540	0.5137	21.1295	703.3440
equity-actions	0.006816	0.0945	21.0788	578.1722
exploration	0.000072	0.0020	41.4920	2033.3655
government	0.000000	0.0000	-	-
health	0.000000	0.0000	69.9500	4894.0002
housing	0.000000	0.0000	-	-
indexes	0.000026	0.0017	68.0168	4664.4898
industrial-accidents	0.000025	0.0015	59.8422	3761.8611
insider-trading	0.000057	0.0036	66.2689	4454.7860
investor-relations	0.000959	0.0258	47.5945	2585.4279
labor-issues	0.003086	0.0578	28.9232	1089.7772
legal	0.000900	0.0258	44.1106	2289.8878
marketing	0.000460	0.0133	56.7864	3563.1334
order-imbalances	0.000000	0.0000	-	-
partnerships	0.001804	0.0393	34.9581	1545.2647
price-targets	0.010833	0.1058	13.7671	256.3751
products-services	0.008895	0.0975	19.9659	574.6577
regulatory	0.000190	0.0094	61.3655	3942.3747
revenues	0.019607	0.2206	22.2481	776.9085
security	0.000018	0.0013	69.9499	4893.9932

stock-picks	0.000081	0.0055	69.2429	4809.5251
stock-prices	0.004652	0.0980	32.3823	1302.9701
taxes	0.000000	0.0000	69.9500	4894.0002
technical-analysis	0.001238	0.0286	36.0347	1537.8846
transportation	0.000012	0.0005	46.8196	2379.6156
war-conflict	0.000021	0.0015	69.9495	4893.9556

B.2. Statistics for cumulated news indicators

Table B.5.: This table shows the statistics for monthly cumulated news sentiment.

Category Group	Mean	Std.Dev.	Skewness	Kurtosis
acquisitions-mergers	4.097870	10.0372	4.8699	34.6602
analyst-ratings	0.285223	2.1442	0.5530	7.5096
assets	0.112004	1.1436	0.2094	30.4094
balance-of-payments	0.000098	0.0249	2.2005	117.5876
bankruptcy	-0.007441	0.0727	-10.6055	134.5159
civil-unrest	-0.001653	0.0196	-10.1976	142.0782
corporate-responsibility	0.015510	0.1040	9.8281	104.2925
credit	0.071114	0.3504	6.9047	64.3838
credit-ratings	-0.110273	1.1551	-1.6464	34.2683
crime	-0.002735	0.0254	-9.8555	126.2079
dividends	0.059792	0.7375	1.2019	46.0359
domestic-product	0.000000	0.0000	-	-
earnings	1.359075	6.6569	0.9789	7.6898
equity-actions	0.652461	2.0830	2.1624	24.0486
exploration	0.045131	0.1328	8.2418	85.0959
government	-0.000562	0.0069	-12.6170	158.8424
health	-0.003076	0.0323	-11.8857	143.7329
housing	0.000000	0.0000	-	-
indexes	0.001973	0.0481	4.1823	135.8303
industrial-accidents	-0.039913	0.3272	-9.8730	106.5639
insider-trading	0.002028	0.1488	3.0315	108.1908
investor-relations	0.000000	0.0000	-	-
labor-issues	0.080718	4.1159	-1.2966	33.3619
legal	-0.746464	3.2589	-5.5225	66.1903
marketing	0.023958	0.1182	8.7498	87.1160
order-imbalances	0.001173	0.1305	2.4355	41.5689
partnerships	1.244608	2.2893	4.5791	30.2547
price-targets	0.076782	0.7814	0.3747	11.6175
products-services	3.397662	5.4941	1.9976	15.3596
regulatory	-0.168496	0.8221	-7.5498	85.7260
revenues	1.051650	2.2533	1.7806	12.0444

security	-0.015238	0.1487	-10.8807	123.4841
stock-picks	0.016339	0.1412	8.3968	100.6677
stock-prices	-0.070841	2.6313	-0.5342	23.7630
taxes	0.000549	0.0056	11.5617	137.6446
technical-analysis	0.110987	1.1249	0.2117	1.9945
transportation	-0.004413	0.0191	-6.2011	42.7930
war-conflict	0.001062	0.0877	-3.4637	130.4666

Table B.6.: This table shows the statistics for monthly cumulated fresh news sentiment.

Category Group	**Mean**	**Std.Dev.**	**Skewness**	**Kurtosis**
acquisitions-mergers	0.709630	1.2771	3.5828	22.4156
analyst-ratings	0.189856	1.1880	0.6237	5.2523
assets	0.035255	0.2380	0.6049	11.7197
balance-of-payments	0.000135	0.0124	3.2053	120.0493
bankruptcy	-0.001484	0.0136	-11.2614	130.5331
civil-unrest	-0.000585	0.0068	-10.2115	141.9772
corporate-responsibility	0.002444	0.0183	9.7263	101.0581
credit	0.017727	0.0890	4.6896	38.7685
credit-ratings	-0.020337	0.3272	-1.2941	21.9634
crime	-0.001157	0.0100	-9.4617	119.9261
dividends	0.052385	0.3373	2.2693	23.0604
domestic-product	0.000000	0.0000	-	-
earnings	0.961549	3.6864	1.1478	9.7771
equity-actions	0.149988	0.5566	1.7361	13.2179
exploration	0.011549	0.0304	7.3720	77.2476
government	-0.000223	0.0028	-12.6496	158.8564
health	-0.000306	0.0033	-11.7625	141.3014
housing	0.000000	0.0000	-	-
indexes	0.000584	0.0191	3.9470	130.8141
industrial-accidents	-0.005097	0.0279	-8.8816	91.1142
insider-trading	0.001158	0.0621	2.0136	82.0423
investor-relations	0.000000	0.0000	-	-
labor-issues	0.084937	0.4537	0.4602	10.4340
legal	-0.086599	0.3076	-3.1308	33.0729
marketing	0.005531	0.0318	8.3066	78.3692
order-imbalances	0.001713	0.0488	2.6524	35.5888
partnerships	0.166598	0.2977	3.0235	14.9864
price-targets	0.060788	0.6215	0.3404	10.7248
products-services	0.593137	0.7689	2.5075	14.6860
regulatory	-0.020664	0.1005	-5.7646	56.4319
revenues	0.469550	1.0328	1.3074	7.2619
security	-0.003134	0.0266	-10.4859	115.2806

stock-picks	0.006773	0.0561	7.5722	86.9212
stock-prices	0.037661	0.5180	0.9961	35.1748
taxes	0.000227	0.0029	12.6573	158.3057
technical-analysis	-0.007125	0.1799	-0.0898	3.3841
transportation	-0.000656	0.0024	-4.6001	24.6788
war-conflict	-0.000607	0.0191	-2.7050	117.7679

Table B.7.: This table shows the statistics for quarterly cumulated news sentiment.

Category Group	Mean	Std.Dev.	Skewness	Kurtosis
acquisitions-mergers	12.209816	19.7127	2.9036	11.2340
analyst-ratings	0.838906	3.9577	0.2960	2.8884
assets	0.332406	2.0940	0.1570	9.8948
balance-of-payments	0.000294	0.0443	1.3938	38.7761
bankruptcy	-0.022269	0.1276	-6.0174	42.1938
civil-unrest	-0.004960	0.0339	-5.7875	44.4759
corporate-responsibility	0.046493	0.1831	5.6634	33.9804
credit	0.212918	0.6152	3.8921	20.0304
credit-ratings	-0.332091	2.2897	-1.1681	12.4822
crime	-0.008205	0.0432	-5.5463	38.8199
dividends	0.177462	1.2890	0.5338	13.6867
domestic-product	0.000000	0.0000	-	-
earnings	4.068246	11.4237	0.0811	1.4586
equity-actions	1.918347	3.8389	1.4285	7.9313
exploration	0.135135	0.2352	4.6973	26.6574
government	-0.001686	0.0122	-7.1686	49.5934
health	-0.009229	0.0587	-7.0619	48.8440
housing	0.000000	0.0000	-	-
indexes	0.005919	0.0840	2.3081	42.8621
industrial-accidents	-0.119634	0.5960	-5.7191	34.3895
insider-trading	0.005891	0.2662	1.6680	39.0770
investor-relations	0.000000	0.0000	-	-
labor-issues	0.226094	7.4086	-0.6966	10.5433
legal	-2.232723	6.2708	-3.2741	21.9794
marketing	0.070470	0.2046	5.0952	28.4687
order-imbalances	0.003520	0.2491	1.2772	14.3608
partnerships	3.695122	4.3421	2.6549	9.5707
price-targets	0.229495	1.6208	0.1768	4.2541
products-services	10.120802	11.3752	1.2965	5.7353
regulatory	-0.499943	1.5917	-4.3938	27.2914
revenues	3.138356	4.0573	0.7244	3.4658
security	-0.045714	0.2598	-6.1477	38.1292
stock-picks	0.048355	0.2487	4.8263	31.9781
stock-prices	-0.213216	4.9721	-0.2491	8.8067

taxes	0.001648	0.0097	6.6150	43.7757
technical-analysis	0.332879	2.4477	0.3452	1.4221
transportation	-0.013238	0.0338	-3.3265	10.6031
war-conflict	0.003185	0.1517	-1.9274	40.3260

Table B.8.: This table shows the statistics for quarterly cumulated fresh news sentiment.

Category Group	**Mean**	**Std.Dev.**	**Skewness**	**Kurtosis**
acquisitions-mergers	2.115705	2.4350	2.1702	7.5178
analyst-ratings	0.564784	2.1453	0.3510	1.9709
assets	0.104429	0.4232	0.3868	4.3939
balance-of-payments	0.000404	0.0215	1.8163	37.1146
bankruptcy	-0.004451	0.0235	-6.3896	40.7502
civil-unrest	-0.001754	0.0117	-5.7652	44.0126
corporate-responsibility	0.007294	0.0315	5.5617	32.4527
credit	0.053014	0.1509	2.5659	11.3745
credit-ratings	-0.061332	0.6075	-0.8874	7.7115
crime	-0.003470	0.0169	-5.3109	36.7549
dividends	0.156233	0.5760	1.0230	5.8014
domestic-product	0.000000	0.0000	-	-
earnings	2.876229	6.0560	0.1293	2.0593
equity-actions	0.434735	0.9883	1.1076	5.4483
exploration	0.034583	0.0524	4.3047	25.2519
government	-0.000668	0.0049	-7.1704	49.5262
health	-0.000919	0.0057	-6.9300	47.2599
housing	0.000000	0.0000	-	-
indexes	0.001751	0.0329	2.2271	40.5276
industrial-accidents	-0.015224	0.0484	-5.1215	29.0695
insider-trading	0.003420	0.1082	1.1549	25.3303
investor-relations	0.000000	0.0000	-	-
labor-issues	0.253288	0.7663	0.2579	3.1670
legal	-0.258820	0.5516	-1.8260	10.6398
marketing	0.016364	0.0535	4.6148	23.5486
order-imbalances	0.005140	0.0772	1.5800	11.1476
partnerships	0.495918	0.5317	1.7515	4.6480
price-targets	0.181679	1.2518	0.1131	3.7699
products-services	1.770690	1.4520	1.4473	4.2118
regulatory	-0.061826	0.1753	-3.2558	17.0245
revenues	1.404859	1.8245	0.2928	1.5836
security	-0.009403	0.0454	-5.8824	35.2023
stock-picks	0.020317	0.0961	4.2224	26.2136
stock-prices	0.111164	0.9208	0.5927	12.3180
taxes	0.000680	0.0050	7.1708	49.4519

technical-analysis	-0.021454	0.2701	-0.1692	0.7167
transportation	-0.001969	0.0044	-3.1411	11.7871
war-conflict	-0.001821	0.0331	-1.4922	36.3270

Table B.9.: This table shows the statistics for weekly cumulated news sentiment.

Category Group	Mean	Std.Dev.	Skewness	Kurtosis
acquisitions-mergers	0.956170	4.2740	9.7580	147.7350
analyst-ratings	0.066552	0.9676	1.0215	27.9097
assets	0.026134	0.5251	0.2130	128.2242
balance-of-payments	0.000023	0.0119	4.4774	518.5430
bankruptcy	-0.001736	0.0335	-22.1973	565.4804
civil-unrest	-0.000386	0.0094	-21.2776	626.0117
corporate-responsibility	0.003619	0.0485	20.3800	456.7523
credit	0.016593	0.1628	13.9483	265.0673
credit-ratings	-0.025730	0.4999	-2.8459	136.8711
crime	-0.000638	0.0123	-20.4360	556.1023
dividends	0.013951	0.3467	2.8337	203.2638
domestic-product	0.000000	0.0000	-	-
earnings	0.317117	3.1326	2.6388	42.6159
equity-actions	0.152241	0.9274	3.9929	100.2026
exploration	0.010531	0.0632	16.8980	358.5111
government	-0.000131	0.0031	-25.3930	656.6415
health	-0.000718	0.0134	-22.7025	545.9220
housing	0.000000	0.0000	-	-
indexes	0.000460	0.0231	8.7226	598.4778
industrial-accidents	-0.009313	0.1515	-20.2311	459.9908
insider-trading	0.000473	0.0698	6.3155	412.8522
investor-relations	0.000000	0.0000	-	-
labor-issues	0.018834	1.9250	-2.6232	143.8382
legal	-0.174175	1.4062	-11.1755	283.0470
marketing	0.005590	0.0563	18.1515	381.6700
order-imbalances	0.000274	0.0581	4.7844	166.2310
partnerships	0.290408	1.0107	9.2529	128.0770
price-targets	0.017916	0.3223	0.8782	43.4711
products-services	0.792788	2.2554	3.7843	55.0886
regulatory	-0.039316	0.3527	-15.2048	361.3256
revenues	0.245385	1.0710	4.3001	59.5589
security	-0.003556	0.0691	-22.6431	543.0296
stock-picks	0.003812	0.0669	17.4684	442.8063
stock-prices	-0.016530	1.1845	-1.1240	97.7980
taxes	0.000128	0.0027	24.1404	607.3087
technical-analysis	0.025897	0.3849	0.0845	3.9907
transportation	-0.001030	0.0091	-13.1238	200.2714

war-conflict	0.000248	0.0412	-7.1407	570.2600

Table B.10.: This table shows the statistics for weekly cumulated fresh news sentiment.

Category Group	**Mean**	**Std.Dev.**	**Skewness**	**Kurtosis**
acquisitions-mergers	0.165580	0.5632	6.9637	92.1483
analyst-ratings	0.044300	0.5457	1.1130	19.1702
assets	0.008226	0.1128	0.9811	48.3147
balance-of-payments	0.000031	0.0060	6.6984	529.0270
bankruptcy	-0.000346	0.0066	-23.1839	559.2539
civil-unrest	-0.000136	0.0033	-21.3174	624.8725
corporate-responsibility	0.000570	0.0089	20.3921	449.1385
credit	0.004136	0.0433	9.9322	177.0325
credit-ratings	-0.004745	0.1508	-2.4000	89.5400
crime	-0.000270	0.0049	-19.5882	528.6284
dividends	0.012223	0.1625	5.0961	108.8128
domestic-product	0.000000	0.0000	-	-
earnings	0.224361	1.7816	3.0665	52.2054
equity-actions	0.034997	0.2389	3.1902	42.6166
exploration	0.002695	0.0149	14.8937	310.2557
government	-0.000052	0.0013	-25.8632	681.9083
health	-0.000071	0.0015	-22.8922	556.7059
housing	0.000000	0.0000	-	-
indexes	0.000136	0.0092	8.2396	575.8753
industrial-accidents	-0.001189	0.0133	-18.3908	399.2886
insider-trading	0.000270	0.0298	4.1047	357.8667
investor-relations	0.000000	0.0000	-	-
labor-issues	0.019819	0.2193	1.1460	44.0163
legal	-0.020206	0.1429	-6.0938	138.0840
marketing	0.001290	0.0155	17.5033	351.4385
order-imbalances	0.000400	0.0246	5.4658	159.5505
partnerships	0.038873	0.1388	5.9775	59.6858
price-targets	0.014184	0.2616	0.8268	38.5472
products-services	0.138399	0.3461	4.9147	61.5165
regulatory	-0.004822	0.0479	-11.8859	245.9292
revenues	0.109562	0.5036	3.4838	40.1271
security	-0.000731	0.0129	-21.9339	510.2216
stock-picks	0.001580	0.0273	15.8331	385.1503
stock-prices	0.008788	0.2451	2.0559	150.9619
taxes	0.000053	0.0014	26.3899	694.8567
technical-analysis	-0.001662	0.0917	-0.2011	18.8088
transportation	-0.000153	0.0011	-8.2932	76.1364

war-conflict	-0.000142	0.0092	-5.7606	514.0412

Bibliography

[AABK15] António Afonso, Michael G. Arghyrou, George Bagdatoglou, and Alexandros Kontonikas. On the time-varying relationship between EMU sovereign spreads and their determinants. *Economic Modelling*, 44: 363 – 371, 2015, http://www.sciencedirect.com/science/article/pii/S0264999314002806.

[ADAWLS18] Shreyash Agrawal, Pablo D. Azar, Andrew W. Lo, and Taranjit Singh. Momentum, mean-reversion, and social media: Evidence from stocktwits and twitter. *The Journal of Portfolio Management*, 44: 85–95, 07 2018, https://doi.org/10.3905/jpm.2018.44.7.085.

[AMS16] David Allen, Michael McAleer, and Abhay Singh. An entropy-based analysis of the relationship between the DOW JONES Index and the TRNA sentiment series. *Applied Economics*, 49: 1–16, 07 2016, https://doi.org/10.1080/00036846.2016.1203067.

[ANG18] Adam Atkins, Mahesan Niranjan, and Enrico Gerding. Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science*, 4, 02 2018, https://doi.org/10.1016/j.jfds.2018.02.002.

[AVEK+13] Cristiano Arbex-Valle, Christina Erlwein, Alexandra Kochendoerfer, Bernhard Kuebler, Gautam Mitra, Giles-Arnaud Nzouankeu-Nana, Barry Nouwt, and Bram Stalknecht. News-enhanced market risk management. *SSRN Electronic Journal*, 01 2013, https://doi.org/10.2139/ssrn.2322668.

[BCP17] Mauro Bernardi, Leopoldo Catania, and Lea Petrella. Are news important to predict the value-at-risk? *The European Journal of Finance*, 23(6): 535–572, 2017, https://doi.org/10.1080/1351847X.2015.1106959.

[BCST17] Renato Bruni, Francesco Cesarone, Andrea Scozzari, and Fabio Tardella. On exact and approximate stochastic dominance strategies for portfolio selection. *European Journal of Operational Research*, 259(1): 322–329, 2017, https://www.sciencedirect.com/science/article/pii/S0377221716308190.

[BFKR13] Jacob Boudoukh, Ronen Feldman, Shimon Kogan, and Matthew P. Richardson. Which news moves stock prices? a textual analysis. Technical Report Working Paper No. w18725, NBER, 2013.

[BGHM17] Jordan Boyd-Graber, Yuening Hu, and David Mimno. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3): 143–296, 2017, http://dx.doi.org/10.1561/1500000030.

[BKLB14] Ilaria Bordino, Nicolas Kourtellis, Nikolay Laptev, and Youssef Billawala. Stock trade volume prediction with Yahoo Finance user browsing behavior. pages 1168–1173, 03 2014, https://doi.org/10.1109/ICDE.2014.6816733.

[BKT20] Salman Baig, Robert Kosowski, and Jerome Teiletche. Newscaster versus nowcaster. *UniGestion*, 09 2020, https://www.unigestion.com/wp-content/uploads/2020/08/Perspectives-Newscasters-vs-Nowcasters-EN-202008.pdf.

[Ble12] David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4): 77–84, April 2012, https://doi.org/10.1145/2133806.2133826.

[BM13] Svetlana Borovkova and Diego Mahakena. News, volatility and jumps: The case of natural gas futures. *SSRN Electronic Journal*, 01 2013, https://ssrn.com/abstract=2334226.

[BO08] Brad M. Barber and Terrance Odean. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies*, 21: 785–818, 02 2008, https://doi.org/10.2139/ssrn.460660.

[Bor15] Svetlana Borovkova. The role of news in commodity markets. *SSRN Electronic Journal*, 03 2015, https://ssrn.com/abstract=2587285.

[BRSOL09] Almira Biglova, Svetlozar T. Rachev, Stoyan Stoyanov, and Sergio Ortobelli Lozza. Analysis of the factors influencing momentum profits. 2009.

[BX15] Svetlana Borovkova and Ding Xiaobo. News sentiment, factor models and abnormal stock returns. *SSRN Electronic Journal*, 11 2015, https://ssrn.com/abstract=2695360.

[Bys16] Hans Byström. Language, news and volatility. *Journal of International Financial Markets, Institutions and Money*, 42: 139–154, 5 2016, https://doi.org/10.1016/j.intfin.2016.03.002.

[CDHH13] Hailiang Chen, Prabuddha De, Yu Hu, and Byoung-Hyoun Hwang. Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies*, 12 2013, https://ssrn.com/abstract=1807265.

[CDK12] Amélie Charles, Olivier Darné, and Jae H. Kim. Exchange-rate return predictability and the adaptive markets hypothesis: Evidence from major foreign exchange rates. *Journal* *of International Money and Finance*, 31(6): 1607–1626, 2012, https://doi.org/10.1016/j.jimonfin.2012.0.

[CFCG09] Beatriz Cuellar, Yolanda Fuertes Callén, and José Antonio Gadea. Stock price reaction to non-financial news in european technology companies. *European Accounting Review*, 20: 81–111, 09 2009, https://doi.org/10.1080/09638180903384650.

[CI18] Dario Caldara and Matteo Iacoviello. Measuring Geopolitical Risk. International Finance Discussion Papers 1222, Board of Governors of the Federal Reserve System (U.S.), February 2018.

[CJRS14] Asher Curtis, Vernon J. Richardson, and Roy Schmardebeck. Investor attention and the pricing of earnings news. *SSRN Electronic Journal*, 01 2014, https://doi.org/10.2139/ssrn.2467243.

[CLM19] Lauren Cohen, Dong Lou, and Christopher J. Malloy. Playing favorites: How firms prevent the revelation of bad news. *SSRN Electronic Journal*, 3 2019, http://dx.doi.org/10.2139/ssrn.2479542.

[Cop15] Zeynep Copur. *Handbook of Research on Behavioral Finance and Investment Strategies: Decision Making in the Financial Industry.* 01 2015.

[Cre15] Germán G. Creamer. Can a corporate network and news sentiment improve portfolio optimization using the black–litterman model? *Quantitative Finance*, 15(8): 1405–1416, 2015, https://doi.org/10.1080/14697688.2015.1039865.

[CSS18] Guglielmo Maria Caporale, Fabio Spagnolo, and Nicola Spagnolo. Macro news and bond yield spreads in the euro area. *The European Journal of Finance*, 24(2): 114–134, 2018, https://doi.org/10.1080/1351847X.2017.1285797.

[DA20] Alexander Denev and Saeed Amen. *The Book of Alternative data: A Guide for Investors, Traders and Risk Managers*. 06 2020.

[DB18] Margot Dijkstra and Svetlana Borovkova. Deep learning prediction of the eurostoxx 50 with news sentiment. 07 2018, https://doi.org/10.13140/RG.2.2.12318.23367.

[Dim18] Ivailo Dimov. Mining news topic codes with sentiment. In *AI, Machine Learing and Sentiment Analysis Applied to Finance*. UNICOM Seminars Ltd, 06 2018.

[DLFM17] Serge Darolles, Gaëlle Le Fol, and Gulten Mero. Mixture of distribution hypothesis: Analyzing daily liquidity frictions and information flows. *Journal of Econometrics*, 201(2): 367 – 383, 2017, http://www.sciencedirect.com/science/article/pii/S0304407617301665.

[DM17] Zhuanxin Ding and R. Douglas Martin. The fundamental law of active management: Redux. *Journal of Empirical Finance*, 43: 91 – 114, 2017, http://www.sciencedirect.com/science/article/pii/S0927539817300543.

[DRT12] Michal Dzielinski, Marc Oliver Rieger, and Tõnn Talpsepp. *Volatility asymmetry, news, and private investors*, chapter 11, pages 255–270. John Wiley & Sons, Ltd, 2012.

[dSPD17] Leonardo dos Santos Pinheiro and Mark Dras. Stock market prediction with deep learning: A character-based neural language model for event-based trading. In *Proceedings of the Australasian Language Technology Association Workshop 2017*, pages 6–15, Brisbane, Australia, 12 2017. https://www.aclweb.org/anthology/U17-1001.

[DZLD15] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Deep learning for event-driven stock prediction. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 2327–2333. AAAI Press, 2015, http://dl.acm.org/citation.cfm?id=2832415.2832572.

[Egi14] Massimo Egidi. The economics of wishful thinking and the adventures of rationality. *Mind & Society*, 13: 9–27, 06 2014, https://doi.org/10.1007/s11299-014-0146-8.

[ES17] Christina Erlwein-Sayer. Forecasting sovereign bond spreads with macroeconomic news sentiment. Technical report, Eurostar - SenRisk Project, 2017.

[ES18a] Christina Erlwein-Sayer. Enhanced prediction of sovereign bond spreads through macroeconomic news sentiment. In *AI, Machine Learing and Sentiment Analysis Applied to Finance*. UNICOM Seminars Ltd, 06 2018.

[ES18b] Christina Erlwein-Sayer. Macroeconomic news sentiment: Enhanced risk assessment for sovereign bonds. *Risks*, 6(4), 2018, http://www.mdpi.com/2227-9091/6/4/141.

[Fam70] Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2): 383–417, 1970, http://www.jstor.org/stable/2325486.

[FMRZ11] Csaba I. Fábián, Gautam Mitra, Diana Roman, and Victor Zverovich. An enhanced model for portfolio choice with ssd criteria: a constructive approach. *Quantitative Finance*, 11(10): 1525–1534, 2011, https://doi.org/10.1080/14697680903493607.

[GB12] UMIT G. GURUN and ALEXANDER W. BUTLER. Don't believe the hype: Local media slant, local advertising, and firm value. *The Journal of Finance*, 67(2): 561–598, 2012, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2012.01725.x.

[GIS17] Robert Charles Giannini, Paul J. Irvine, and Tao Shu. Nonlocal disadvantage: An examination of social media sentiment. *SSRN Electronic Journal*, 04 2017, https://ssrn.com/abstract=2958998.

[GKG+14] Maxim Gusev, Dimitri Kroujiline, Boris Govorkov, Sergey V Sharov, Dmitry Ushanov, and Maxim Zhilyaev. Sell the news? A news-driven model of the stock market. Technical Report arXiv:1404.7364, 04 2014.

[GKG+15] Maxim Gusev, Dimitri Kroujiline, Boris Govorkov, Sergey V. Sharov, Dmitry Ushanov, and Maxim Zhilyaev. Predictable markets? A news-driven model of the stock market. MPRA Paper 58831, University Library of Munich, Germany, 2015.

[GKW19] Qi Ge, Alexander Kurov, and Marketa Halova Wolfe. Do investors care about presidential company-specific tweets? *Journal of Financial Research*, 2019, https://onlinelibrary.wiley.com/doi/abs/10.1111/jfir.12177.

[Gol16] Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 11(57): 345–420, 2016, https://doi.org/10.1613/jair.4992.

[Haf16] Peter Hafez. Exploiting alternative data in the investment process. In *4th RavenPack Research Symposium - Reshaping Finance with alternative data*. RavenPack, 2016, https://youtu.be/Hc42mxHEGqU.

[Haf18] Peter Hafez. News sentiment everywhere! In *AI, Machine Learing and Sentiment Analysis Applied to Finance*. UNICOM Seminars Ltd, 06 2018.

[HGC15] Peter Hafez and Jose A. Guerrero-Colòn. Using sentiment to create theme-based alphas. *RavenPack Quantitative Research*, 09 2015.

[HGCD15] Peter Hafez, Jose A. Guerrero-Colòn, and Stefan Duprey. Thematic alpha streams improve performance of equity portfolios. *RavenPack Quantitative Research*, 09 2015.

[HGCG+19] Peter Hafez, Jose A. Guerrero-Colòn, Francisco Gomez, Maria Gomez, and Ricard Matas. Fundamental investing: Big data, big biotech. *RavenPack Quantitative Research*, 01 2019.

[HKGM19] Peter Hafez, Marko Kangrga, Francisco Gomez, and Ricard Matas. Deutsche bank: Where is the marginal investor? *RavenPack Quantitative Research*, 04 2019.

[HL69] G. Hanoch and H. Levy. The Efficiency Analysis of Choices Involving Risk. *The Review of Economic Studies*, 36(3): 335–346, 07 1969, https://doi.org/10.2307/2296431.

[HLGC+18a] Peter Hafez, Francesco Lautizi, Jose A. Guerrero-Colòn, Francisco Gomez, Maria Gomez, and Ricard Matas. Citigroup's take on capex announcement strategies using ravenpack data. *RavenPack Quantitative Research*, 07 2018.

[HLGC+18b] Peter Hafez, Francesco Lautizi, Jose A. Guerrero-Colòn, Francisco Gomez, Maria Gomez, and Ricard Matas. Effects of event sentiment aggregation: Sum vs. mean. *RavenPack Quantitative Research*, 05 2018.

[HLGC+18c] Peter Hafez, Francesco Lautizi, Jose A. Guerrero-Colòn, Francisco Gomez, Maria Gomez, and Ricard Matas. Empirical research partners' take on big data: Media sentiment - useful around the world. *RavenPack Quantitative Research*, 10 2018.

[HLGC+18d] Peter Hafez, Francesco Lautizi, Jose A. Guerrero-Colòn, Francisco Gomez, Maria Gomez, and Ricard Matas. J.p. morgan: Using ravenpack sentiment for cross asset style timing. *RavenPack Quantitative Research*, 10 2018.

[HLGC+18e] Peter Hafez, Francesco Lautizi, Jose A. Guerrero-Colòn, Francisco Gomez, and Ricard Matas. How intraday news patterns dictate the implementation of sentiment strategies. *RavenPack Quantitative Research*, 08 2018.

[HLGC+19] Peter Hafez, Francesco Lautizi, Jose A. Guerrero-Colòn, Francisco Gomez, and Ricard Matas. A news sentiment stock screener for discretionary traders. *RavenPack Quantitative Research*, 05 2019.

[HLS+19] Peter Hafez, Francesco Lautizi, Duprey Stefan, Jose A. Guerrero-Colòn, and Mads Koefoed. Achieve high capacity strategies trading economically-linked companies. *RavenPack Quantitative Research*, 05 2019.

[HM13] Matthew Hull and Frank Mcgroarty. Do emerging markets become more efficient as they develop? long memory persistence in equity indices. *Emerging Markets Review*, 18, 11 2013, https://doi.org/10.1016/j.ememar.2013.11.001.

[HMG+20] Peter Hafez, Ricard Matas, Inna Grinis, Francisco Gomez, Marko Kangrga, Boris Skorodumov, and Alan Liu. Trading around the earnings calendar. *RavenPack Quantitative Research*, 09 2020.

[HML+19] Peter Hafez, Ricard Matas, Francesco Lautizi, Francisco Gomez, and Jose A. Guerrero-Colòn. A news sentiment stock screener for discretionary traders. *RavenPack Quantitative Research*, 05 2019.

[HN17] Andreas Haupenthal and Matthias Neuenkirch. Grexit news and stock returns. *Applied Economics*, 49(39): 3891–3898, 2017, https://doi.org/10.1080/00036846.2016.1270418.

[Hoc15] Ronald Hochreiter. Computing trading strategies based on financial sentiment data using evolutionary optimization. In *MENDEL*, 2015.

[HR69] Josef Hadar and William R. Russell. Rules for ordering uncertain prospects. *The American Economic Review*, 59(1): 25–34, 1969, http://www.jstor.org/stable/1811090.

[HS16] Steven L. Heston and Nitish R. Sinha. News versus sentiment: Predicting stock returns from news stories. *Finance and Economics Discussion Series*, 2016: 1–35, 06 2016, https://ssrn.com/abstract=2792559.

[HS17] Thanh D. Huynh and Daniel R. Smith. Stock price reaction to news: The joint effect of tone and attention on momentum. *Journal of Behavioral Finance*, 18(3): 304–328, 2017, https://doi.org/10.1080/15427560.2017.1339190.

[IKSSS16] Feriha Ibriyamova, Samuel Kogan, Galla Salganik-Shoshan, and David Stolin. Using semantic fingerprinting in finance. *Applied Economics*, 49: 1–17, 11 2016, https://ssrn.com/abstract=2755585.

[KD12] Petko S. Kalev and Huu Nhan Duong. *Firm-Specific news Arrival and the Volatility of Intraday Stock Index and Futures Returns*, pages 271–288. 05 2012.

[Key36] John Maynard Keynes. *The State of Long-Term Expectation*, chapter 12. Palgrave Macmillan, 1936.

[KGU+15] Dimitri Kroujiline, Maxim Gusev, Dmitry Ushanov, Sergey V. Sharov, and Boris Govorkov. Forecasting stock market returns over multiple time horizons. *SSRN Electronic Journal*, 08 2015, https://doi.org/10.2139/ssrn.2646909.

[Kit12] John Kittrell. *Sentiment Reversals as buy Signals*, pages 231–244. 05 2012.

[KJF17] Ahmet Karagozoglu and Frank J Fabozzi. Volatility wisdom of social media crowds. *The Journal of Portfolio Management*, 01 2017, https://doi.org/10.3905/jpm.2017.43.2.136.

[KKX20] Zheng Ke, Bryan T. Kelly, and Dacheng Xiu. Predicting returns with text data. *SSRN Electronic Journal*, 9 2020, http://dx.doi.org/10.2139/ssrn.3389884.

[KPN07] Imre Kondor, Szilárd Pafka, and Gábor Nagy. Noise sensitivity of portfolio selection under various risk measures. *Journal of Banking & Finance*, 31(5): 1545–1573, 2007, https://www.sciencedirect.com/science/article/pii/S0378426607000052.

[Kri13] Ladislav Kristoufek. Can Google Trends search queries contribute to risk diversification? *Scientific reports*, 3: 2713, 09 2013, https://doi.org/10.1038/srep02713.

[KSL11] Jae H. Kim, Abul Shamsuddin, and Kian-Ping Lim. Stock return predictability and the adaptive markets hypothesis: Evidence from century long U.S. data. *Journal of Empirical Finance*, 18: 868–879, 12 2011, https://doi.org/10.1016/j.jempfin.2011.08.002.

[Lev90] Haim Levy. *Stochastic dominance*. The Macmillan Press, 1990.

[Lev92] Haim Levy. Stochastic dominance and expected utility: Survey and analysis. *Management Science*, 38(4): 555–593, 1992, http://www.jstor.org/stable/2632436.

[Liu18] Huicheng Liu. Leveraging financial news for stock trend prediction with attention-based recurrent neural network. 11 2018, https://arxiv.org/abs/1811.06173.

[Lo05] Andrew W. Lo. Reconciling efficient markets with behavioral finance : The adaptive markets hypothesis. *Journal of Investment Consulting*, 7(2): 21–44, 2005, https://ssrn.com/abstract=1702447.

[LQLW17] Yifan Liu, Zengchang Qin, Pengyu Li, and Tao Wan. Stock volatility prediction using recurrent neural networks with sentiment analysis. In *IEA/AIE*, 2017.

[LRW+18] Y. Luo, G. Rohal, S. Wang, M. Alvarez, J. Jussa, and J. Zhong. Banking on the banks – welcome to BALI. *Wolfe Research Luo's QES*, 09 2018, http://www.wolferesearch.com/research-library/x20180904_YL_MDA_Banks.pdf.

[LRW+19] Y. Luo, G. Rohal, S. Wang, M. Alvarez, J. Jussa, and J. Zhong. Beyond fake news. *Wolfe Research Luo's QES*, 01 2019.

[LXB18] Jia Li, Yuan Xue, and Tim Bollerslev. Volume, Volatility, and Public News Announcements. *The Review of Economic Studies*, 85(4): 2005–2041, 01 2018, https://doi.org/10.1093/restud/rdy003.

[Mar59] Harry M. Markowitz. *Portfolio Selection: Efficient Diversification of Investments*. Yale University Press, 1959.

[MCA+13] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Kenett, H. Eugene Stanley, and Tobias Preis. Quantifying Wikipedia usage patterns before stock market moves. *Scientific reports*, 3: 1801, 05 2013, https://doi.org/10.1038/srep01801.

[MdBBY15] Gautam Mitra, Dan di Bartolomeo, Ashok Banerjee, and Xiang Yu. Automated analysis of news to compute market sentiment: Its impact on liquidity and trading. *SSRN Electronic Journal*, 05 2015, https://ssrn.com/abstract=2605049.

[MESVY18] Gautam Mitra, Christina Erlwein-Sayer, Cristiano Valle, and Xiang Yu. *Using Market Sentiment to Enhance Second-Order Stochastic Dominance Trading Models*, chapter 2, pages 25–48. CRC Press, 02 2018.

[Meu10] Attilio Meucci. The Black-Litterman approach: Original model and extensions. *ARPM - Advanced Risk and Portfolio Management*, 10 2010, https://dx.doi.org/10.2139/ssrn.1117574.

[Meu11] Attilio Meucci. A new breed of copulas for risk and portfolio management. *Risk*, 24: 122–126, 05 2011, https://ssrn.com/abstract=1752702.

[Min92] Hyman P. Minsky. The financial instability hypothesis. 1992, http://dx.doi.org/10.2139/ssrn.161024.

[ML03] Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 188–191, 2003, https://www.aclweb.org/anthology/W03-0430.

[MMD08] Leela R. Mitra, Gautam Mitra, and Dan Dibartolomeo. Equity portfolio risk (volatility) estimation using market information and sentiment. *Quantitative Finance*, 9: 887–895, 12 2008, https://doi.org/10.2139/ssrn.1425624.

[Moe14] Richhild Moessner. Government bond yield sensitivity to economic news at the zero lower bound in Canada in comparison with the UK and US. *Applied Financial Economics*, 24(11): 739–751, 2014, https://doi.org/10.1080/09603107.2014.902019.

[MOL17] Matteo Malavasi and Prof. Sergio Ortobelli Lozza. Semiparametric tests for behavioral finance efficiency. In *Financial Management of Firms and Financial Institutions: 11th International Scientific Conference, Proceedings (Part II)*. VŠB-TU of Ostrava, Faculty of Economics, Department of Finance, 09 2017, http://hdl.handle.net/10446/122796.

[MOLT19] Matteo Malavasi, Prof. Sergio Ortobelli Lozza, and Prof. Stefan Trueck. *Essays on Stochastic Orderings in Portfolio Selection*. Ph.d. dissertation, University of Bergamo, Bergamo, Italy, 2019.

[MOT21] Matteo Malavasi, Sergio Ortobelli Lozza, and Stefan Trück. Second order of stochastic dominance efficiency vs mean variance efficiency. *European Journal of Operational Research*, 290(3): 1192–1206, 2021, https://www.sciencedirect.com/science/article/pii/S0377221720307645.

[MS99] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. 05 1999.

[MV18] Lorenzo Malandri and Prof. Carlo Vercellis. *Investigating the Impact of Public Mood in the Stock Market*. Ph.d. dissertation, Politecnico di Milano, Milano, Italy, 2018.

[MX16] G. Mitra and Y. Xiang. *Handbook of Sentiment Analysis in Finance.* Albury Books, 2016.

[NM13] Inna Novalija and Dunja Mladenic. Applying semantic technology to business news analysis. *Applied Artificial Intelligence*, 27: 520–550, 2013, https://doi.org/10.1080/08839514.2013.805600.

[NN17] Seema Narayan and Paresh Kumar Narayan. Are oil price news headlines statistically and economically significant for investors? *Journal of Behavioral Finance*, 18(3): 258–270, 2017, https://doi.org/10.1080/15427560.2017.1308942.

[OSOL15] Rapheal Olaniyan, Daniel Stamate, Lahcen Ouarbya, and Doina Logofătu. Sentiment and stock market volatility predictive modelling — a hybrid approach. pages 1–10, 10 2015, https://doi.org/10.1109/DSAA.2015.7344855.

[OT15] Sergio Ortobelli and Tomáš Tichý. On the impact of semidefinite positive correlation measures in portfolio theory. *Annals of Operations Research*, 236: 625–652, 2015, https://doi.org/10.1007/s10479-015-1962-x.

[PAFKN+14] Matija Piškorec, Nino Antulov-Fantulin, Petra Kralj Novak, Igor Mozetič, Miha Grčar, Irena Vodenska, and Tomislav Smuc. Cohesiveness in financial news and its relation to market volatility. *Scientific reports*, 4: 5038, 05 2014.

[PPNK05] Gabor Papp, Szilard Pafka, Maciej A. Nowak, and Imre Kondor. Random Matrix Filtering in Portfolio Optimization. 09 2005, https://arxiv.org/abs/physics/0509235.

[Rav17] RavenPack. *RavenPack Analytics User Guide and Service Overview Version 1.0.* RavenPack, 2017.

[RBB+14] Gabriele Ranco, Ilaria Bordino, Giacomo Bormetti, Guido Caldarelli, Fabrizio Lillo, and Michele Treccani. Coupling news sentiment with web browsing data predicts intra-day stock prices. 12 2014, http://dx.doi.org/10.2139/ssrn.2699167.

[RC15] Gabriele Ranco and Prof. Guido Caldarelli. *Predictive power of web Big Data in Financial Economics.* Ph.d. dissertation, IMT Institute for Advanced Studies, Lucca, Italy, 2015.

[Reu12] Reuters. *Thomson Reuters News Analytics Data Message Guide Version 2.1.1.* Thomson Reuters, 2012.

[RLJW18] G. Rohal, Y. Luo, J. Jussa, and S. Wang. Global TMT stock selection models. *Wolfe Research Luo's QES*, 11 2018, http://www.wolferesearch.com/research-library/x20181106_YL_MDA_TMT.pdf.

[RLW17] G. Rohal, J. Luo, Y. amd Jussa, and S. Wang. Text mining unstructured corporate filing data. *Wolfe Research Luo's QES*, 04 2017, http://wolferesearch.com/research-library/x20170420_YL_MDA_QR.pdf.

[RMF⁺07] Svetlozar T. Rachev, Stefan Mittnik, Frank J. Fabozzi, Sergio M. Focardi, and Teo Jaić. *Principal Component Analysis and Factor Analysis*, chapter 13. Wiley, 2007.

[RMF⁺15a] Svetlozar T. Rachev, Stefan Mittnik, Frank J. Fabozzi, Sergio M. Focardi, and Teo Jasic. *Cointegration and State Space Models*, chapter 11, pages 373–405. John Wiley & Sons, Ltd, 2015.

[RMF⁺15b] Svetlozar T. Rachev, Stefan Mittnik, Frank J. Fabozzi, Sergio M. Focardi, and Teo Jasic. *Principal Components Analysis and Factor Analysis*, chapter 13, pages 429–464. John Wiley & Sons, Ltd, 2015.

[RMZ11] Diana Roman, Gautam Mitra, and Victor Zverovich. Enhanced indexation based on second-order stochastic dominance. *European Journal of Operational Research*, 228, 03 2011, https://ssrn.com/abstract=1776966.

[Ros76] Stephen A Ross. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3): 341–360, 1976, https://www.sciencedirect.com/science/article/pii/0022053176900466.

[Ros78] Stephen A Ross. Mutual fund separation in financial theory—the separating distributions. *Journal of Economic Theory*, 17(2): 254–286, 1978, https://www.sciencedirect.com/science/article/pii/002205317890073X.

[Row90] Thomas Harvey Rowan. *Functional stability analysis of numerical algorithms*. PhD thesis, Department of Computer Sciences, University of Texas at Austin, 1990.

[RR09] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado, June 2009. Association for Computational Linguistics, https://www.aclweb.org/anthology/W09-1119.

[Rut13] A. Ruttiens. Portfolio risk measures: The time's arrow matters. *Computational Economics*, 41, 03 2013.

[SDM18] Zryan A Sadik, Paresh M Date, and Gautam Mitra. News augmented GARCH(1,1) model for volatility prediction. *IMA Journal of Management Mathematics*, 30(2): 165–185, 03 2018, https://doi.org/10.1093/imaman/dpy004.

[SDM19] Zryan Sadik, Paresh Date, and Gautam Mitra. Forecasting crude oil futures prices using global macroeconomic news sentiment. *IMA Journal of Management Mathematics*, 06 2019.

[Sha66] William F. Sharpe. Mutual fund performance. *The Journal of Business*, 39(1): 119–138, 1966, http://www.jstor.org/stable/2351741.

[Sha94] William F. Sharpe. The sharpe ratio. *The Journal of Portfolio Management*, 21(1): 49–58, 1994, https://jpm.pm-research.com/content/21/1/49.

[SKB15] Semenov Semen, Dr. Drona Kandhai, and Dr. Svetlana Borovkova. News and behavioral impact on commodity markets through multi-agent modeling. Msc dissertation, University of Amsterdam, 2015.

[SKTA15] Thársis Souza, Olga Kolchyna, Philip Treleaven, and Tomaso Aste. Twitter sentiment analysis applied to finance: A case study in the retail industry. 07 2015, https://arxiv.org/abs/1507.00784.

[Sma16] Lee A. Smales. Time-varying relationship of news sentiment, implied volatility and stock returns. *Applied Economics*, 48(51): 4942–4960, 2016, https://doi.org/10.1080/00036846.2016.1167830.

[SMS19] Zryan Sadik, Gautam Mitra, and Berry Shradha. Asset allocation strategies: Enhanced by micro-blog. *SSRN Electronic Journal*, 12 2019, https://ssrn.com/abstract=3527510.

[SMT20] Zryan Sadik, Gautam Mitra, and Ziwen Tan. Asset allocation strategies: Enhanced by news. *SSRN Electronic Journal*, 04 2020, https://ssrn.com/abstract=3588364.

[SVM14] Carlos Sorzano, Javier Vargas, and A. Montano. A survey of dimensionality reduction techniques. 03 2014.

[SW11] Timm O. Sprenger and Isabell Welpe. News or noise? the stock market reaction to different types of company-specific news events. *SSRN Electronic Journal*, 01 2011, https://ssrn.com/abstract=1734632.

[SYY17] Jia-Lang Seng, Pi-Hua Yang, and Hsiao-Fang Yang. Initial public offering and financial news. *Journal of Information and Telecommunication*, 1(3): 259–272, 2017, https://doi.org/10.1080/24751839.2017.1347762.

[TCK16] Daniel Tsvetanov, Jerry Coakley, and Neil Kellard. Is news related to GDP growth a risk factor for commodity futures returns? *Quantitative Finance*, 16(12): 1887–1899, 2016, https://doi.org/10.1080/14697688.2016.1211797.

[Tet07] Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3): 1139–1168, 2007, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2007.01232.x.

[Tha15] Eric Tham. The unbearable lightness of expectations of the Chinese investor. *Handbook of Sentiment Analysis in Finance*, 11 2015, https://ssrn.com/abstract=3168041.

[Tob58] James Tobin. Liquidity Preference as Behavior Towards Risk. *Review of Economic Studies*, 25(2): 65–86, 1958, https://ideas.repec.org/a/oup/restud/v25y1958i2p65-86..html.

[Tob65] James Tobin. *The theory of portfolio selection*. Macmillan, 1965.

[UH13] Andrew Urquhart and Robert Hudson. Efficient or adaptive markets? evidence from major stock markets using very long historic data. *International Review of Financial Analysis*, 28: 130–142, 06 2013, https://doi.org/10.1016/j.irfa.2013.03.005.

[UM14] Andrew Urquhart and Frank Mcgroarty. Calendar effects, market conditions and the adaptive market hypothesis: Evidence from long-run U.S. data. *International Review of Financial Analysis*, 35, 08 2014, https://doi.org/10.1016/j.irfa.2014.08.003.

[UM16] Andrew Urquhart and Frank Mcgroarty. Are stock markets really efficient? evidence of the adaptive market hypothesis. *International Review of Financial Analysis*, 47, 07 2016, https://doi.org/10.1016/j.irfa.2016.06.011.

[UPM15] Matthias W Uhl, Mads Pedersen, and Oliver Malitius. What's in the news? using news sentiment momentum for tactical asset allocation. *The Journal of Portfolio Management*, 41: 100–112, 12 2015, https://ssrn.com/abstract=2490011.

[WKM12] Dariusz Wojcik, Nicholas Kreston, and Sarah McGill. Freshwater, saltwater, and deepwater: Efficient market hypothesis versus behavioral finance. *Journal of Economic Geography*, 13, 02 2012, https://doi.org/10.2139/ssrn.2008788.

[XSLL13] Yunqing Xia, Weifeng Su, Raymond Y.K. Lau, and Yi Liu. Discovering latent commercial networks from online financial news articles. *Enterprise Information Systems*, 7(3): 303–331, 2013, https://doi.org/10.1080/17517575.2011.621093.

[YI15] John Robert Yaros and Tomasz Imieliński. Data-driven methods for equity similarity prediction. *Quantitative Finance*, 15, 10 2015, https://doi.org/10.1080/14697688.2015.1071079.

[YM18] Xiang Yu and Gautam Mitra. Enhanced trading strategy using sentiment and technical indicators. In *AI, Machine Learing and Sentiment Analysis Applied to Finance*. UNICOM Seminars Ltd, 06 2018.

[YMAVS15] Xiang Yu, Gautam Mitra, Cristiano Arbex-Valle, and Tilman Sayer. An impact measure for news: Its use in daily trading strategies. *SSRN Electronic Journal*, 01 2015, https://ssrn.com/abstract=2702032.

[YML15] Steve Y. Yang, Sheung Yin Kevin Mo, and Anqi Liu. Twitter financial community sentiment and its predictive relationship to stock market movement. *Quantitative Finance*, 15(10): 1637–1656, 2015, https://doi.org/10.1080/14697688.2015.1071078.

[YMY13] Xiang Yu, Gautam Mitra, and Keming Yu. Impact of news on asset behaviour: Return, volatility and liquidity in an intra-day setting. *SSRN Electronic Journal*, 07 2013, https://ssrn.com/abstract=22968552.

[ZESM17] Cai Zhixin, Christina Erlwein-Sayer, and Gautam Mitra. Enhanced corporate bond yield modelling incorporating macroeconomic news sentiment. Technical report, Eurostar - SenRisk Project, 2017.

[ZHCB16] Junni L. Zhang, Wolfgang K. Härdle, Cathy Y. Chen, and Elisabeth Bommes. Distillation of news flow into analysis of stock reactions. *Journal of Business & Economic Statistics*, 34(4): 547–563, 2016, https://doi.org/10.1080/07350015.2015.1110525.