ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D MTEC

Angelo D'Anna

Disentangling idiosyncratic and cross-sectional market sentiment from news reports to predict stock price movements

Master Thesis

Chair of Entrepreneurial Risks
Swiss Federal Institute of Technology (ETH) Zurich

Supervision

Prof. Dr. Didier Sornette
Prof. Dr. Sandro Lera

December 2021

Abstract

All possible types of news influence the stock market. In this study, we seek to disentangle market sentiment from news and only retain the idiosyncratic sentiment in order to better predict next-day stock price movements. To achieve this goal, we propose two approaches to extract firm-specific information from news articles. In the first approach, we study the syntax of the sentences containing firm-specific information and we retain only the ones where a firm is either the subject or the object of the sentence. In the second approach, we use sentence embedding to calculate the similarity between firm-specific and market information and retain only those with low similarity. These two approaches are broadly comparable, and both can be used to predict the price of a particular stock. We use FinBERT, a bidirectional encoder representations from transformers (BERT) model pre-trained on large financial text corpora, to extract the sentiment of the news which we use as trading signal for our sentiment-based investment strategy. Lastly, we use a large set of news records from Thompson Reuters to backtest our investment strategy and assess our performance using idiosyncratic sentiment compared to headline or news body sentiment. Our results show significantly better investment performance using the sentiment of the idiosyncratic parts of news than using the sentiment of news headlines and bodies.

Contents

1 Introduction **1**

2 Related Work **3**

2.1 Using News to Predict Stock Price Movements 3

2.2 Previous Work on Firm-specific News 3

3 Data **5**

3.1 News Data 5

3.2 Price and Market Capitalization Data 5

4 Methodology **7**

4.1 Data Cleansing and Pre-processing 7

4.2 Disentangling the Idiosyncratic Part of News 7

4.2.1 Syntax Approach 8

4.2.2 Embedding Approach 9

4.3 Signal Extraction 11

4.4 Investment Strategy 13

4.5 Evaluation Metrics 14

5 Results **16**

5.1 Comparison between Idiosyncratic Approaches 16

5.2 Investment Strategy Results 18

5.3 Limitations 20

6 Conclusions **22**

Bibliography **22**

A BERT and FinBERT **26**

B Additional Results **28**

B.1 Results with Neutral News 28

B.2 Kolmogorov-Smirnov two-sample Tests Results 29

C Additional Robustness Checks: Idiosyncratic Approaches **34**

1 Introduction

From the time the first share was traded on a stock exchange, every professional or private investor has been interested in predicting stock price movements.

The random walk theory states that consecutive changes in the price of a security are independent, identically distributed random variables, which implies that past events cannot be used to predict future price movements. The independence of successive price changes is consistent with the hypothesis of "efficient" markets, which implies that at every point in time, given the available information, the actual price of a security reflects a good estimate of its intrinsic value (Fama, 1965); therefore, the stock price reflects all the available information, and it moves in response to news and events.

Investment decisions highly depend on subjective and objective factors (Virlics, 2013). Thanks to the invention of the internet, accessing and exchanging information has become easier and faster, allowing investors to make decisions based on a wider spectrum of information. With the evolution of computing power, new ways of processing information have been explored and applied to both structured and unstructured data. As a result, machine learning (ML), textual analysis, and natural language processing (NLP) have increasingly been employed to predict how stock prices will move based on financial news.

Different studies have shown that applying NLP techniques to news articles makes it possible to understand the articles' tone and sentiment and predict the stock market's behaviour (Tetlock, 2007; Narayan & Bannigidadmath, 2017; Li et al., 2020). In most studies, models are trained using headlines (Ding et al., 2015; Deng et al., 2019; D. Chen et al., 2019; Li et al., 2020) from news articles or the body of the news (Tetlock, 2007; Ke et al., 2019) to predict price movements, but at the time this study was conducted little research had been undertaken on how prediction models perform when using different parts of news articles, and in particular on predicting stock price movements using firm-specific parts of news.

In Figure 1.1, we show the headline and body of an article, published by Reuters on May 3, 2012[1]. Not all the relevant information can be contained in the semantic of the headline. In our study, we test the following hypothesis: selecting only firm-specific parts of the news (i.e. the idiosyncratic component of the news) and using their sentiment for a sentiment-based investment strategy can lead to better investment performances compared to using the sentiment of the headlines or the entire body of the news articles.

In this study, we analyse how the performance of a sentiment-based investment strategy changes using different news components, and we test if removing the *market component* and retaining only the *idiosyncratic component* of the news can lead to better investment results. In particular, we propose two methods to separate firm-specific information from the rest of the article; we use FinBERT[2], a fine-tuned BERT model, to classify the sentiment of the idiosyncratic parts, the headlines, and the entire body of news, and use those sentiment scores as a trading signal for our investment strategy. We backtest these strategies on a set of more than 200'000 news records from Reuters[3] using daily closing price data on more than 1000 stocks, included in the MSCI World Index[4].

In Chapter 2, we present an overview of recent studies on stock market movements prediction using NLP applied to news and on the use of firm-specific news to study market behaviours. In Chapter 3, we present our dataset of news, stock closing prices, and market capitalisation data. In Chapter 4, we explain how we prepare the data for our analysis, the methods we use to extract the idiosyncratic part of the news, how we create the sentiment scores that we will use for our

[1] https://www.reuters.com/article/bmw-idUSL5E8G30XO20120503
[2] https://github.com/yya518/FinBERT
[3] https://www.reuters.com
[4] https://www.msci.com/World

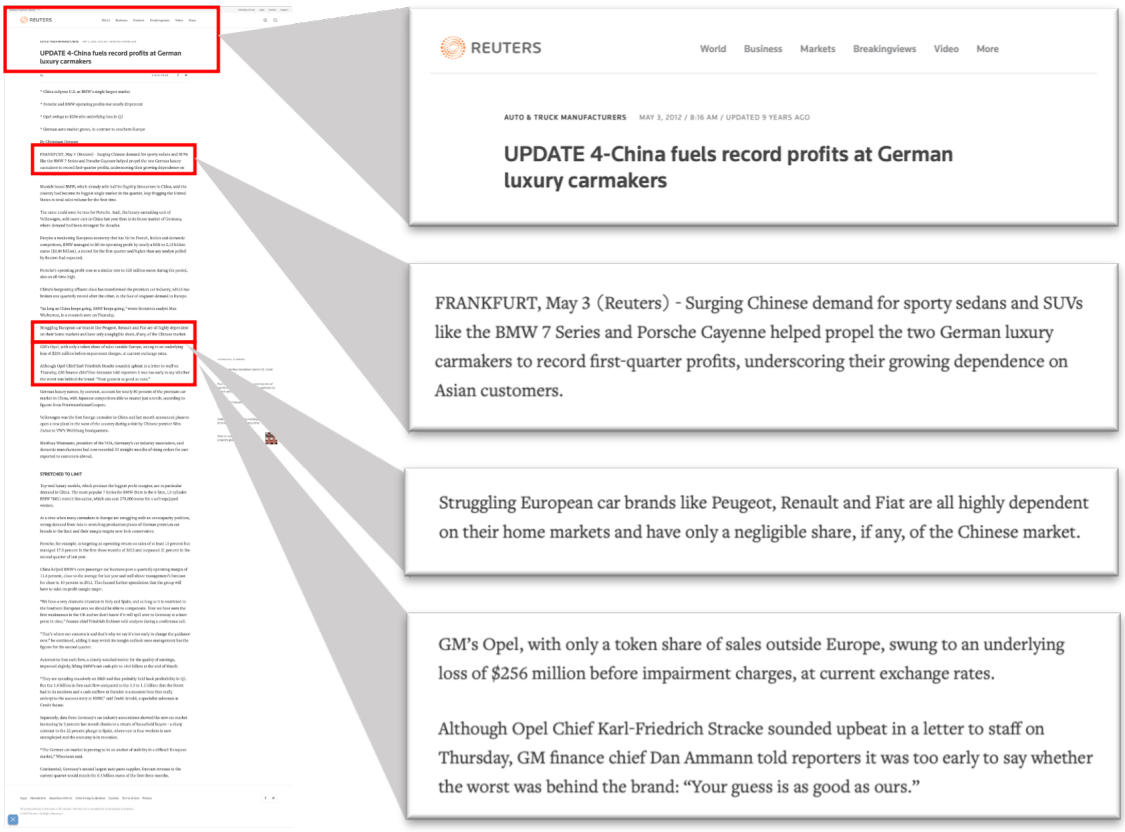

Figure 1.1: Article from Reuters from May 3 2012.

investment strategy, and the metrics we employ to evaluate our investment strategies. In Chapter 5, we present the results of our analysis, together with the limitations of our model and some inputs for future works. Finally, in Chapter 6, we make our conclusions.

2 Related Work

In the last decades, thanks to advancements in computing, NLP and machine learning, it has been easier to analyze unstructured data like text. These advancements have made it easier to investigate the relation between news and stock prices. For this reason, various studies have been conducted trying to find the best stock price prediction model, using state-of-the-art NLP and ML techniques. In the following sections, we show recent studies in the field of predicting stock price movements using news and previous work on firm-specif news.

2.1 Using News to Predict Stock Price Movements

Hui et al. (2017) study whether the positioning of the text can affect the reader's sentiment toward the news article and differentiate between three parts of the text: headline, first paragraph, and last paragraph. The headline usually summarizes the content of the entire article and is more important for categorizing the reader's feelings. The first paragraph usually describes the main idea of the whole article and plays an equally important role as the headline in classifying news reports by their sentiments. The last paragraph typically gives a detailed description of the event, even though it does not play a significant role in classifying the reader's sentiment as the other parts of the article. Consistent with this study, Ding et al. (2014) find that using only news headlines to predict stock price movements achieves better performances compared to using only the body of news or a combination of news body and headlines. According to the authors, one reason may be the extraction of irrelevant information from news bodies. The majority of studies (Ding et al., 2015; Deng et al., 2019; D. Chen et al., 2019; Li et al., 2020) use the same approach as Ding et al. (2014) and use only the headlines of news to predict stock price movements.

Tetlock (2007) is the first to find proofs that news articles can be used to predict stock market activity. Using the General Inquirer's Harvard IV-4 psychological dictionary to classify whether a word is positive or negative, he creates a measure of media pessimism and applies it on a famous column of the Wall Street Journal[1], finding that high media pessimism predicts high trading volumes and bearish movements on market prices. More recent studies apply deep learning methods to predict stock market movements. Ding et al. (2014, 2015) use structured events and convolutional neural networks on news headlines to predict individual stocks and the S&P 500, Deng et al. (2019) use a knowledge-driven temporal convolutional network on news from Reddit WorldNews Channel[2] combined with Dow Jones Industrial Average price values to forecast stock trends. Ke et al. (2019) use a different approach and follow a three steps model where they first isolate a set of sentiment terms through predictive screening, they attribute sentiment weights to these terms via topic modelling, and finally, they create an article-level sentiment score via penalized likelihood. Li et al. (2020) use both technical indicators and sentiment from news articles to feed a neural network able to make stock predictions. They use a finance domain-specific sentiment dictionary (Loughran-McDonald Financial Dictionary) to extract news sentiment.

2.2 Previous Work on Firm-specific News

The goal of all the papers cited in the previous section is to predict the price movements of indexes or sets of stocks based either on the headlines or on the entire body of the news, without addressing the problem of trying to separate market-specific information from firm-specific information.

Ryan & Taffler (2004) study the relationship between capital market information flows and changes in company share prices and trading volume activity. The authors examine the relative

[1] https://www.wsj.com/
[2] https://www.reddit.com/r/worldnews/

importance of firm-specific information events in driving corporate price changes and trading volume. They use a manual matching process to classify news items into information categories. The study finds that reported corporate news events are responsible for the majority of economically significant price changes and volume fluctuations for companies.

DeLisle et al. (2016) use firm-specific news (i.e. announcement or declaration of share repurchases, debt issuances, seasoned equity offerings, merger and acquisition targets, M&A acquirers, insider trades, analyst recommendations, earnings, dividends and stock splits) to study the relationship between idiosyncratic volatility and stock returns around news releases. They do not focus on the news text but rather on the date the news was published. Contrary to the limited arbitrage explanation for the negative price of idiosyncratic volatility, they find that the idiosyncratic volatility unrelated to news announcements is strongly negatively priced. In contrast, the idiosyncratic volatility related to firm-specific news announcements is positively priced.

Engle et al. (2021) also investigate the relation between firm-specific news and volatility, showing that changes in return volatility can be partially explained by public information arrival. In their work, they use a sample of 28 large U.S. companies included in the Dow Jones Industrial Average and take all the news items identified by the Dow Jones Intelligent Indexing system as being related to a particular firm. Also, in this case, they do not look at the text of the news, but they use the Dow Jones Intelligent Indexing categorization to distinguish between the content of news items (e.g. Bankruptcy, Dividends, etc.), and they use a simple count of news items for each news category for their analysis.

At the time this study was conducted, no major research had been done on how to disentangle firm-specif information from news articles and use it as a trading signal. This study wants to give a new perspective on the relevance and potential of disentangling firm-specific information from news and contribute to the always increasing literature on stock price movements predictions.

3 Data

This study is conducted on a large set of news items published between 2007 and 2018. From our dataset, we extract different parts of news, and we analyse them throughout the study. We calculate the sentiment score of each news part and use it as an investment signal for our sentiment-based investment strategy. In the end, we backtest our investment strategies and assess their performances on historical data using our panel data on daily closing prices for more than 1000 stocks.

3.1 News Data

The news data-set used for this study includes 202,435 news items from Reuters[1] published between 2007-01-18 and 2018-09-22.

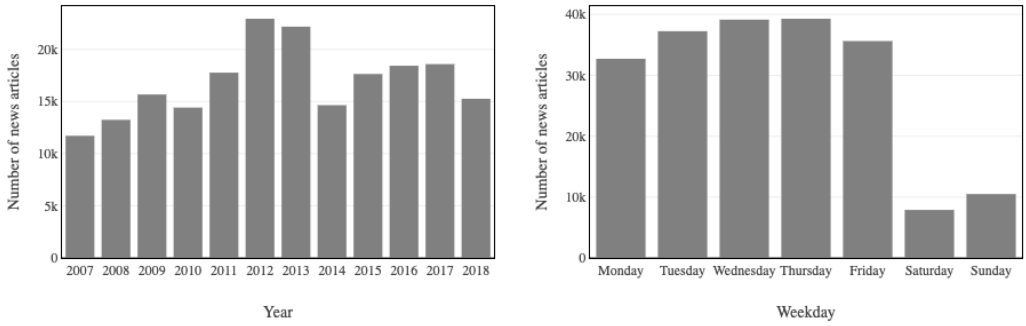

Figure 3.1: The plot on the left shows the number of news in our dataset per year. The plot on the right shows the number of news in our dataset per weekday.

In Figure 3.1, we can see the number of news items per year and per weekday. Each entry in our dataset contains a date stamp, a headline, a description (equal to the first sentence of the body), a body, and sometimes a location. Having both the bodies and the headlines of the articles is fundamental for our study since it allows us to make a comparison between the two. A sample entry of the dataset is presented in Table 3.1.

Since the *location_ id* is not relevant to our study and the *description* is just redundant, we drop these two columns from our dataset. To identify the companies mentioned in the articles and link them to their share prices, we extract the Reuters instrument codes (RICs) from the news bodies. The RICs are ticker-like combinations of alphanumerical characters used to distinguish financial instruments or indices. The RICs consist of a ticker symbol, a period and the stock exchange code of the ticker. For example, "JPM.N" is a valid RIC which refers to JP Morgan Chase & Co.'s[2] stock traded on the New York Stock Exchange. "JPM.L" instead refers to the same stock trading on the London Stock Exchange. This last exchange code used in the RIC is proprietary to Thomson Reuters.

3.2 Price and Market Capitalization Data

The prices panel data used in our analysis include daily closing prices information on 1394 stocks part of the MSCI World Index from 1997-05-30 until 2021-05-26. Look-ahead bias refers to using

[1] https://www.reuters.com
[2] https://www.jpmorgan.com/

Table 3.1: Sample entry of the dataset.

date_published	2012-03-14 21:30:55
location_id	nan
headline	Treasury to sell stock in six bailed-out banks.
description	The U.S. Treasury said on Wednesday that it plans to sell its preferred stock position in six community banks as part of the Obama administration's effort to unwind bailout programs from the financial crisis. Treasury
body	WASHINGTON (Reuters) - The U.S. Treasury said on Wednesday that it plans to sell its preferred stock position in six community banks as part of the Obama administration's effort to unwind bailout programs from the financial crisis. Treasury said it plans to conduct public auctions to sell its stock in Banner Corp (BANR.O), First Financial Holdings Inc FFCH.O, MainSource Financial Group MSFG.O, Seacoast Banking Corp (SBCF.O), Wilshire Bancorp WIBC.O and WSFS Financial Corp (WSFS.O). It has so far recovered $259 billion from the Troubled Asset Relief Program bank programs, which were set up to help stabilize the financial system during the 2007-09 crisis. The Treasury said it expects to start unwinding its positions in the six small banks around March 26 using a modified Dutch auction that establishes a market price by allowing investors to submit bids at specified increments. Treasury still has stakes in 361 banks. Last week the administration announced plans to cut its position in insurer American International Group (AIG.N) to 70 per cent from 77 per cent.

information that would not be available during the period being simulated, which usually results in an upward shift in the results. Survivorship bias refers to the fact that many investment performance estimates are based on datasets that include only funds that existed at the end of the sample period (Daniel et al., 2009). We avoid survivorship and look-ahead bias by checking every day of our analysis if that stock is part of the MSCI World Index at that time, if not, we exclude that stock from our analysis for that day.

Using the panel data on the market capitalization of the companies, we create a capitalization-weighted and an equal-weighted index and compare them to the actual MSCI World Index. Further in our study, we will use the capitalization-weighted index as a benchmark for our investment strategies.

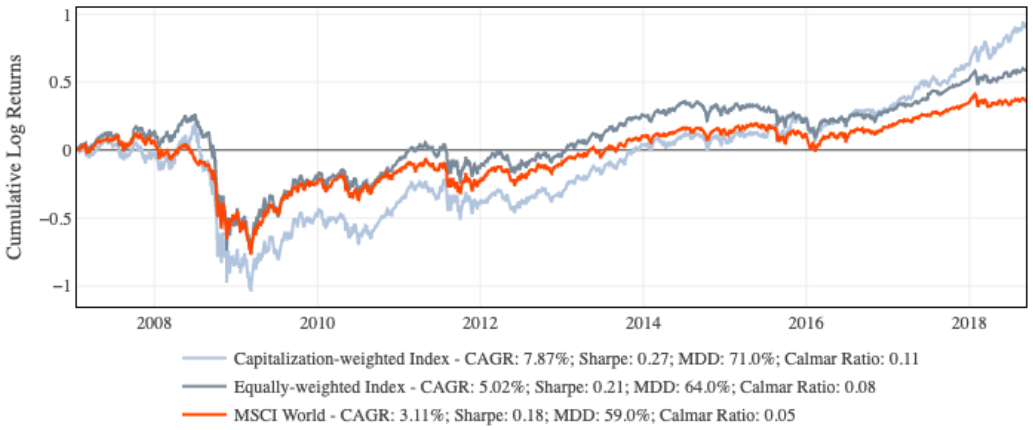

Figure 3.2: Comparison of the cumulative log-returns of a buy-and-hold strategy between 2007-01-18 and 2018-09-22 for the MSCI World Index, the capitalization-weighted index, and the equally-weighted index we created. The legend shows the average cumulative annual growth rate (CAGR), the annualized Sharpe ratio, the maximum drawdown (MDD) and the Calmar ratio of each index.

In Figure 3.2, we compare the performance of a buy-and-hold strategy of the MSCI World Index to the performance of the capitalization-weighted index and the equally-weighted index we created during the period between 2007-01-18 and 2018-09-22.

4 Methodology

Q. Chen (2021) states that an article about a specific stock with a positive sentiment is a predictor of a rise in the stock price of the company, while news articles with negative sentiments are predictors of negative performance for a stock, and news with neutral sentiments do not give any information on the stock's price movement (therefore, investors would not base their investment decision on those). Most studies focus on finding the state of the art ML model to predict stock price movements using textual data from tweets or news. In our study, using already consolidated techniques, we propose two methods to disentangle idiosyncratic pieces of information from news data, and we use them to create a trading signal.

The intuition behind our study is that news articles have both a *market component* and an *idiosyncratic component*. The *market component* contains general information on market developments, while the *idiosyncratic component* only contains firm-specific information. We want to test if using only the sentiment of the firm-specific parts of the news (i.e. the idiosyncratic part) as a trading signal for our investment decisions leads to higher returns and better risk profiles than using the sentiment of the entire body of the articles or only the headlines. In our study, we follow three steps to prepare the news data for our investment strategy: data cleansing, separation of the idiosyncratic part of the news from the rest of the body, and extraction of the sentiment.

4.1 Data Cleansing and Pre-processing

Using unstructured data like news requires preliminary cleaning and pre-processing to discard irrelevant information. Except for the embedding approach that we will introduce in the next sections, we filter out from our dataset all the news articles where no RIC is mentioned in the body because we need to satisfy one fundamental requirement: we need to link the articles' content to the stocks mentioned within the news. In the absence of a RIC, it is not a trivial task to link firm names to their tickers: it could lead to errors in associating the sentiment of the news to the correct stock symbol, given that in the data set, we have hundreds of different firms mentioned. This way, our data set of news is reduced from 202,435 news items to roughly 75,000 news items.

To process the news headlines, we do not do any additional cleansing or pre-processing because they already come as a separate attribute in our dataset.

Our first step to process the news bodies is to remove unnecessary text parts like "Washington (Reuters) -" or "By Lauren Tara La Capra (Reuters) -" that do not give any additional information to the article. As a second step, we split the entire text into sentences because FinBERT, the model we use to extract the sentiment from the text, cannot process more than 512 tokens at a time. The sentiment score of the body of an article will therefore be the arithmetic mean of the sentiment scores of the single sentences contained in the body of the news.

To process the idiosyncratic part of the news, we initially follow the same steps as for the body of the news; as a third step, we only retain the firm-specific part of the text. In the following sections, we introduce the methods we used to extract the idiosyncratic parts of news.

4.2 Disentangling the Idiosyncratic Part of News

In our study, we define the *idiosyncratic part* of the news as the part of the article that contains firm-specific information. To extract only the sentences containing information about a specific company from the news body, we need to define a method to distinguish between the *market component* and the *idiosyncratic component* of news. For example, if we consider the body of the news in Table 3.1, the first sentence:

> "WASHINGTON (Reuters) - The U.S. Treasury said on Wednesday that it plans to sell its preferred stock position in six community banks as part of the Obama administration's effort to unwind bailout programs from the financial crisis."

does not contain any firm-specific information but only provides information on certain market developments. On the contrary, the second sentence contains firm-specif information:

> "Treasury said it plans to conduct public auctions to sell its stock in Banner Corp (BANR.O), First Financial Holdings Inc FFCH.O, MainSource Financial Group MSFG.O, Seacoast Banking Corp (SBCF.O), Wilshire Bancorp WIBC.O and WSFS Financial Corp (WSFS.O)".

Our task is to isolate the idiosyncratic parts of news and decide whether to invest in the companies mentioned in these news articles or not. To make this connection between the companies mentioned in the news and their stock prices, we need to have a clear and straightforward reference to the stock ticker linked to the company. To cope with this requirement, we select only the sentences of the news articles where at least one RIC is mentioned.

4.2.1 Syntax Approach

After having retained only the sentence with at least one RIC, we want to make sure that the information in the sentence pertains to the company mentioned; therefore, we apply an additional filter to our subset of sentences, and we retain only the sentences where a company is either the subject or the object of the phrase. To check if the subject or object of a sentence is an organisation, we use spaCy[1], an open-source library for NLP in Python, which can process and "understand" large volumes of text. In particular, thanks to its part-of-speech tagging and named entities recognition features, it can recognise both the syntactic dependency of words in a sentence (i.e. understand which token is the subject, the verb, the object, etc. of a sentence) and recognise real-world objects such as organisations, dates, locations, persons etc. and denote them with a proper name.

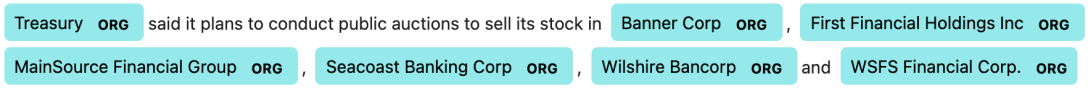

Figure 4.1: Example of spaCy's named entities recognition feature.

Thanks to the named entities features, we can identify any organisation mentioned in the sentence to later check if this organisation is either the subject or the object of the phrase. In Figure 4.1, we can see an example of how spaCy's named entity feature works.

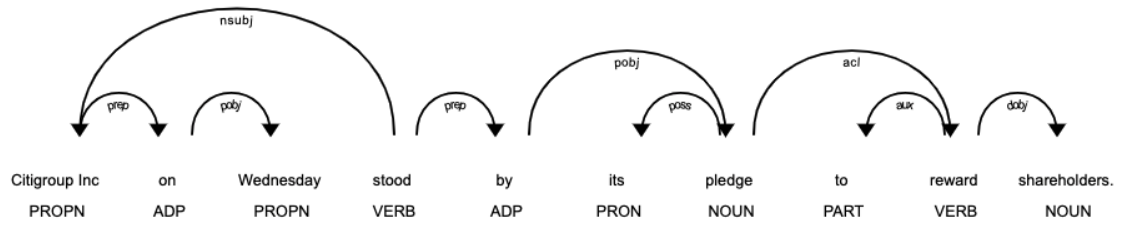

Figure 4.2: Example of spaCy's dependencies parser feature.

[1] https://spacy.io

Once we distinguish which noun chunk[2] in the sentence represents an organisation, we use the dependency parser from spaCy to check the syntactic dependency between the organisation and the rest of the phrase. In Figure 4.2, we show an example of how spaCy's dependency parser works. For our definition of idiosyncratic sentences, we take all the sentences where the organisation is either the subject or the object of the sentence. In particular, we use spaCy's "merge_entity" and "merge_noun_chunks" pipeline components which respectively merge named entities and noun chunks into single tokens (e.g. the sentence "Goldman Sach's share price went up last night" instead of being split into single tokens like "Goldman", "Sach", "'s", "share", "price", "went", "up", "last", "night", will be split into named entities and noun chunks like "Goldman Sach's share price", "went", "up", "last night") and if the token containing the organisation name is either the nominal subject, the passive nominal subject, the direct object, or the object predicative (J. D. Choi et al., 2016) we will consider the sentence as idiosyncratic. In Table 4.1, we show some examples of sentences that, according to this approach, are classified as idiosyncratic part of news or not.

Table 4.1: Examples of sentences that, according to our syntax approach, are classified as idiosyncratic parts of news or not.

Sentence	Idiosyncratic
Nissan Motor Co. Ltd.'s (7201.T) sales rose 3.9 percent, driven by a big increase in its luxury Infiniti division, while Honda Motor Co. Ltd. (7267.T) sales were up 7.3 percent.	Yes
American International Group Inc (AIG.N) on Thursday reported a 17 percent fall in quarterly profit as its general insurance business failed to show improvement, missing analysts' expectations.	Yes
Orchard Supply Hardware Stores Corp OSH.O, spun off by Sears Holdings Corp (SHLD.O) less than two years ago, has filed for Chapter 11 bankruptcy protection partly blaming hefty dividends paid out to its former parent.	Yes
The German sportswear firm, which has been losing ground for years to fast-growing rival Nike (NKE.N), said it was testing automated production units that would allow it to shift manufacturing from Asia closer to consumers.	No
Chairman Kaspar Villiger has said Ermotti – already being groomed as a possible successor since he joined UBS from UniCredit SpA (CRDI.MI) in April – was a strong candidate to take over as CEO permanently.	No

4.2.2 Embedding Approach

The second method we are going to test in this study is more experimental and uses sentence embeddings (Reimers & Gurevych, 2019). Instead of checking the syntax of the sentences with at least one RIC, in this approach, we look at the semantic of the sentences and filter out the clauses that contain market information.

Each day we collect all the news published within a specific period Δt[3], and we select only the news where no RICs are mentioned in the body and where the headline does not contain any company name. In this subset of news, we expect a bigger concentration of market information. To recognise if a company name is contained in the news headline, we use the named entity recognition feature from spaCy. Since spaCy considers entities like the "FED", the "EU", and other political or governmental entities as organizations, we create a subset of entities that will not be considered as companies[4]. Once we have selected the subset of news containing market information, we split each news body into single sentences and filter out the sentences where at least a company name is mentioned (once again, we use spaCy to identify company names). This way, we ensure that our *market component* set of news does not contain any firm-specific information. We finally convert

[2]Nouns chunks are nouns followed by the words describing them, for example, "insurance liabilities" or "promising quarter".

[3]As we explain in the next section, we use a Δt equal to the time between the last closing time of the NYSE and 20 minutes before the next closing.

[4]We choose a subset of organization names from the 50 most mentioned organizations in our set of news: "FED", "EU", "ECB", "Treasury", "Federal Reserve", "OPEC", "IMF", "European Union", "European Central Bank", "SEC", "Congress", "International Monetary Fund", "European Commission", "Securities and Exchange Commission".

each sentence in this set into meaningful sentence embeddings that can be compared using cosine-similarity. Reimers & Gurevych (2019) introduce the concept of sentence embedding: they add a pooling operation to the output of BERT to derive a fixed-sized sentence embedding. They experiment using three pooling strategies: using the hidden state of the "[CLS]" token; calculating the mean of all the output vectors (mean pooling); and calculating a max-over-time of the output vectors. To convert the sentences into sentence embeddings, in this study we use *all-mpnet-base-v2*[5], a model based on MPNet (Song et al., 2020) which has been pre-trained on a large dataset of over one billion training pairs and uses mean pooling to create sentence embeddings, mapping each sentence to a 768-dimensional dense vector space.

Once we have our daily *market component* set of sentences $M_{\Delta t}$, we can focus on isolating the idiosyncratic part of news published within the period Δt . We start by selecting the news articles which are not part of the *market component* set of news, and for each news we select only the sentences where at least one RIC is mentioned. Subsequently, if a sentence is compound, meaning that it is made up of two or more independent clauses joined by a coordinating conjunction or by a comma or semicolon, we split the sentences into single clauses; for example, the sentence

> "In the U.S., most major banks passed a Fed stress test with flying colours and JPMorgan's (JPM.N) dividend hike bolstered investor confidence in the banking sector."

will be split into two clauses: "In the U.S. most major banks passed a Fed stress test with flying colours, and", and "JPMorgan 's (JPM.N) dividend hike bolstered investor confidence in the banking sector". To split sentences into clauses, we use spaCy's dependencies parser. We identify the root[6] of the sentence and its conjunctions, and if the conjunctions belong to different branches of the three, we separate the branches of the three into single clauses. In Figure 4.3 we show the tree structure for the sentence we used as an example.

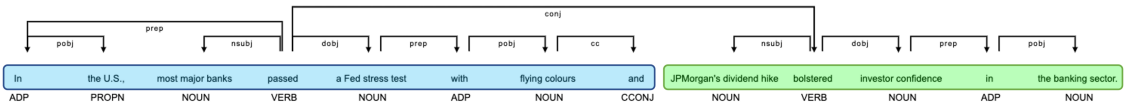

Figure 4.3: Example of tree structure of a sentence. The root of the sentence, in this case, is the verb "passed". The tree has two main branches (one for each clause).

Each clause in our set of clauses $\Upsilon_{\Delta t}$ will be converted into a sentence embedding. In (4.1) we define the cosine distance $\cos(\mathbf{A}, \mathbf{B})$ between two generic vectors $\mathbf{A}$ and $\mathbf{B}$.

$$\cos(\mathbf{A}, \mathbf{B}) := \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \tag{4.1}$$

For each clause i published in Δt we calculate the minimum, the median, and the mean cosine similarity (4.2) between the dense vector $\mathbf{i}$ of the clause $i \in \Upsilon_{\Delta t}$ and the dense vector of each sentence contained in the daily *market component* dataset $M_{\Delta t}$.

$$\begin{aligned}
&C(\mathbf{i})_{M_{\Delta t}} = \{\cos(\mathbf{i}, \mathbf{j}) \mid j \in M_{\Delta t}\} \ , \ i \in \Upsilon_{\Delta t} \\
&C^{\min}_{\Delta t} = \{c^{\min}_{\mathbf{i}} | c^{\min}_{\mathbf{i}} = \min\ C(\mathbf{i})_{M_{\Delta t}}, \ i \in \Upsilon_{\Delta t}\} \\
&C^{\text{median}}_{\Delta t} = \{c^{\text{median}}_{\mathbf{i}} \ | c^{\text{median}}_{\mathbf{i}} = \text{median}\ C(\mathbf{i})_{M_{\Delta t}}, \ i \in \Upsilon_{\Delta t}\} \\
&C^{\text{mean}}_{\Delta t} = \{c^{\text{mean}}_{\mathbf{i}} \ | c^{\text{mean}}_{\mathbf{i}} = \overline{C(\mathbf{i})}_{M_{\Delta t}}, \ i \in \Upsilon_{\Delta t}\}
\end{aligned} \tag{4.2}$$

In (4.3) we show that the set of idiosyncratic clauses $I(m)_{\Delta t}$ that we use for our investment strategy on a day t is made of all the clauses i published in the time span Δt for which $c_{\mathbf{i}}$ is bigger

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[6] The root of a sentence is the head of the entire structure, it does not depend on any node in the dependencies tree.

than the m^{th} percentile $P_{m\Delta t}$ of the cosine distance set $C_{\Delta t}$. With $C_{\Delta t}$ we refer to one of the three cosine distance sets presented in (4.2) (i.e. $C_{\Delta t}^{\text{min}}$, $C_{\Delta t}^{\text{median}}$, and $C_{\Delta t}^{\text{mean}}$) and with $c_{\mathbf{i}}$ we refer to the respective measures (i.e. $c_{\mathbf{i}}^{\text{min}}$, $c_{\mathbf{i}}^{\text{median}}$, and $c_{\mathbf{i}}^{\text{mean}}$).

$$I(m)_{\Delta t} = \{i | c_{\mathbf{i}} > P_{m\Delta t}\} \tag{4.3}$$

In the results chapter, we will show how different measures and different percentiles influence the results. In Table 4.2, we show some examples of similarities between clauses and market component.

Table 4.2: Examples of similarities between clauses and the *market component* set of news. For each clause, we show the most similar sentence belonging to the market component in the same period Δt when the clause was published and the similarity scores (the smaller the cosine distance, the more similar). If the sentence is compounded (i.e. consists of multiple clauses) we show the rest of the compounded sentence in italics into squared brackets.

Clause	Market Component	$c_{\mathbf{i}}^{\text{min}}$	$c_{\mathbf{i}}^{\text{median}}$	$c_{\mathbf{i}}^{\text{mean}}$
Data showing U.S. consumer sentiment rose to its highest level in more than four years in early May , but [*concerns over Europe and JPMorgan Chase & Co's (JPM.N) $2 billion trading loss led equity markets to retreat.*]	U.S. consumer sentiment rose to its highest level in more than four years in early May as Americans were upbeat about the job market and buying plans improved, a survey showed on Friday, offering an encouraging sign for the economic recovery.	0.128	0.850	0.839
U.S. stock and bond markets will be closed on Tuesday, but [*the two-largest U.S. stock exchange operators, NYSE Euronext NYX.N and Nasdaq OMX Group (NDAQ.O), intend to reopen Wednesday, conditions permitting.*]	U.S. stock and options markets will be closed on Monday, and possibly Tuesday, as regulators, exchanges and brokers worry about the integrity of markets and the safety of employees in the face of Hurricane Sandy.	0.172	0.947	0.932
The interest-rate-sensitive plays, including Citigroup Inc (C.N) and JPMorgan Chase & Co (JPM.N), also got a lift from comments from Federal Reserve Chairman Ben Bernanke, who said late on Thursday a resurgence in financial strains in recent weeks had dimmed the outlook for the U.S economy, raising speculation that policy-makers are willing to lower benchmark rates again.	Federal Reserve Chairman Ben Bernanke said on Thursday a resurgence in financial strains in recent weeks had dimmed the outlook for the U.S. economy, signaling an openness to lowering interest rates again.	0.165	0.645	0.681
Earnings on Thursday for major banks JP Morgan (JPM.N), Citigroup (C.N) and Wells Fargo beat expectations (WFC.N), but [*each showed evidence of slower loan growth.*]	The enforcement action follows a recent consumer backlash against the company and its senior management over a series of revelations about its corporate culture and business tactics, including complaints of sexual harassment.	0.865	0.945	0.961

4.3 Signal Extraction

Textual data, particularly news, is a good source of information that allows for predicting stock price movements. For example, a news item about a firm that includes terms or phrases like "disappointing" or "downgrade" might give an investor the signal that there could be a drop in the stock price of that firm (Chan & Franklin, 2011). Early studies on news sensitive stock price movements prediction have used shallow features extractions methods. Schumaker & Chen (2009) use bag-of-words (BoW), noun phrases and named entities to represent text from financial news articles. BoW uses the occurrence of each word as a feature to represent sentences disregarding grammar and words order. Noun phrasing instead uses only a subset of terms as features; using a lexicon, nouns are identified within sentences and aggregated through syntactic rules on the rest of the text, creating noun phrases. Finally, named entities use lexical semantic/syntactic tagging and classify nouns and noun phrases under predefined categories (e.g. dates, locations, organizations, persons, etc.). Shallow feature extraction techniques' main draw-down is that they cannot capture information in the form of events or structured entities relations. For this reason, shallow features

are not the most accurate in representing the impact of news events on stock prices movements predictions (Ding et al., 2014). However, structured representation of events increases sparsity and therefore decreases the predictive power of the features (Ding et al., 2015). To cope with sparsity, discrete and sparse elements like words can be represented as continuous vectors through embedding, low-dimensional, learned continuous vector representations of text, where words with the same semantic have similar representations.

The bidirectional encoder representations from transformers (BERT) model introduced by Devlin et al. (2018) is a language model commonly used for NLP applications which has been pre-trained on large data (WikiBooks[7] and Wikipedia[8]) to learn the basic features of communication. BERT is a trained transformer encoder stack with a large number of transformer blocks (or encoder layers). It comes in two versions, the "Base" version, with twelve transformer blocks, and the "Large" version, with 24 transformer blocks.

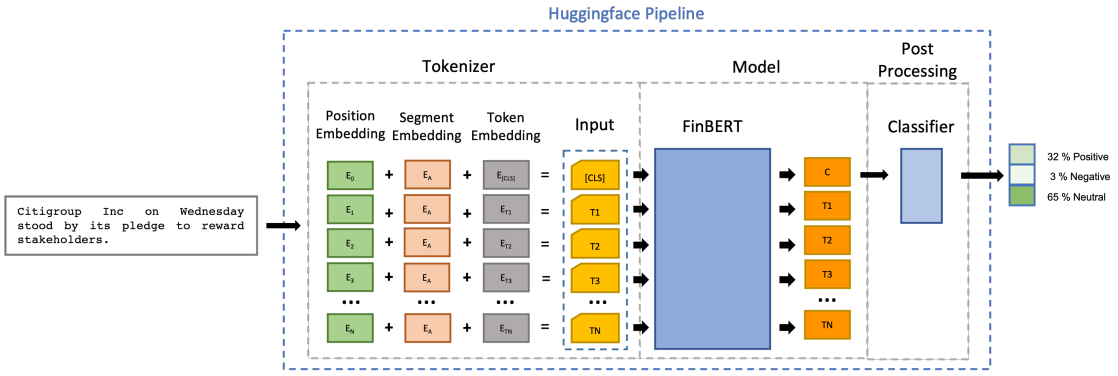

Figure 4.4: Structure of the *pipeline()* function for sentiment analysis in Huggingface. The pipeline consists of three stages: tokeniser, model, and post-processing. Giving raw text as input to the pipeline, this is converted into input IDs which are the sum of a position embedding (indicating the position of each token in the raw text), the segment embedding (which, in case of two sentences packed together, it indicates to which sentence the token belongs to), and the token embedding (created using WordPiece embedding) (Wu et al., 2016). Those inputs IDs are fed to the model, which outputs logits. Those logits are transformed into labels and scores through the post-processing step using a SoftMax layer.

FinBERT is a finance-specific language BERT model pre-trained on large financial text corpora of 4.9 billion tokens, including earning conference call transcripts, corporate reports, and analyst reports (for comparison, BERT's pre-training corpora consists of a total of 3.3 billion tokens) (Yang et al., 2020). In our study, we use the pre-trained FinBERT model (Araci, 2019) available on the Huggingface model repository[9] (Wolf et al., 2019). In Figure 4.4 we show and explain the functioning of the *pipeline()* function for sentiment analysis in Huggingface with FinBERT; for more details on the functioning of BERT and FinBERT refer to Appendix A. The model provides softmax outputs for each text given as input for three labels: positive, negative and neutral. Qiao et al. (2019) illustrate how BERT allocates learned attention to different types of tokes. They distinguish between three groups of tokens: markers (part of this group are the tokens "[CLS]" and "[SEP]", respectively put at the beginning and the end of the sentence given as input to the model, the final hidden state of the first token represents the aggregated sentence and is used for classification tasks, while the "[SEP]" token is used to separate sentence pairs and is used to predict if, given two sentences as input, the second one naturally follows the first one; Devlin & Chang 2018), stop-words and regular words. They find that, while the markers receive the most attention, the stop-words received as much attention as non-stop words, and removing stop-words has no impact on the mean reciprocal rank (MRR)[10] performances. For this reason, we decide to not remove stop-words from the sentences that we feed to FinBERT.

[7] http://en.wikibooks.org/
[8] http://en.wikipedia.org/
[9] https://huggingface.co/ProsusAI/finbert
[10] The MRR is a measure used to evaluate processes that, given a set of queries as inputs, give as output a list

Araci (2019) shows how FinBERT fails in classifying the right sentiment label mainly in two cases:

Case 1: when numerical data are provided in the absence of words indicating directions like "increased".
Example: "Losses totalled USD 0.1 million, compared to previous year USD 2.5 million losses.". True value: Positive. Predicted: Negative.

Case 2: when a statement indicates polarity about a company or the amount of information is not enough.
Example: "It is very important for the bank to pass this year's stress test". True value: Neutral. Predicted: Positive.

In Table 4.3, we show some examples of sentiments extracted using FinBERT.

Table 4.3: Examples of sentiment scores using FinBERT. P_+, P_-, and $P_=$ are respectively the probabilities that the sentence belongs to the positive, negative or neutral class given by the SoftMax classification layer.

Sentence	P_+	P_-	$P_=$
Closing arguments in the U.S. trial of former HSBC Holdings Plc (HSBA.L) executive Mark Johnson came to an end Wednesday, with a lawyer for the government urging jurors to convict Johnson of defrauding a client.	0.02	0.89	0.09
For young children, there's now insurance for recalled infant milk formula, and for little ones who get out of hand People's Insurance Group of China Co Ltd (PICC) (1339.HK) offers a policy against "mischievous and destructive" habits.	0.08	0.03	0.89
Britain's biggest retailer Tesco (TSCO.L) recorded its worst quarterly UK sales drop in 40 years, raising questions over boss Phil Clarke's strategy to counter the challenges of a rapidly-changing grocery industry.	0.01	0.97	0.02
Boeing Co's (BA.N) board raised the company dividend about 50 percent on Monday and approved $10 billion in new share buyback authority that the company said it would use in the next two to three years.	0.81	0.01	0.08

4.4 Investment Strategy

In an ideal world, investors would read all the news published during the day, classify them into positive, negative or neutral, assign them a score, and eventually make an investment decision based on their sentiment on the future price movement of the stocks. However, due to the massive volume of news published daily, it is not realistic to react to all news because not all information significantly move the market and transaction fees would hardly make it profitable. A more logical approach would consist of reading all the news published within a particular period and investing only in the stocks that will more likely have a change in price based on the sentiment of the news. Neutral news will generally be ignored because they are not expected to move the market significantly (Q. Chen, 2021). In Appendix B, we show how the results change when not filtering out neutral news.

In our setting, we want to replicate the behaviour of an investor who uses as a trading signal the sentiment of news, therefore every day we only use the most positive and negative news to rebalance our portfolio. As done by Q. Chen (2021), we define the news sentiment score $S(news)$ as:

$$S(news) = (P_+(news) - 0.5) * 2 \tag{4.4}$$

where *news* from now on will indicate one of the three parts of the news on which we base our

of possible responses, ordered by the probability of correctness. The reciprocal rank is defined as the multiplicative inverse of the position of the correct answer in the list of answers. The MRR is simply the average of the reciprocal ranks of a sample of queries.

analysis: headline, body, idiosyncratic part. With $P_+(news)$[11] we denote the positive sentiment score of the news (i.e. the probability that a sentence belongs to the positive class given by the classification model).

We use $P_{n\Delta t}$ to indicate the n^{th} percentile of all the sentiment scores $S(news)_{\Delta t}$ of news published within a period Δt. The set of news on which we will trade can be defined as $E(n)_{\Delta t}$ where:

$$\begin{aligned} E(n)^-_{\Delta t} &= \{news_{\Delta t} | S(news_{\Delta t}) < P_{n\Delta t}\} \\ E(n)^+_{\Delta t} &= \{news_{\Delta t} | S(news_{\Delta t}) > P_{100-n\Delta t}\} \\ E(n)_{\Delta t} &= E(n)^-_{\Delta t} \cup E(n)^+_{\Delta t} \end{aligned} \tag{4.5}$$

In the strategy we will adopt for our study, we use Δt equal to the time between the last closing time of the NYSE and 20 minutes before the next closing time. This means that every day the market is open, we collect all the news published between the last market closure and twenty minutes before the market closes, we calculate the sentiment score of all the news published in Δt, we get our trading news set $E(n)_{\Delta t}$, and we buy all the stocks mentioned in the news set $E(n)^+_{\Delta t}$ and short the stocks mentioned in $E(n)^-_{\Delta t}$ at the closing price, rebalancing our portfolio. For example, on Monday at 3:40 p.m. ET, we would calculate the sentiment score $S(news)$ of each news published between last Friday at 4:00 p.m. ET and Monday at 3:40 p.m. ET, and trade on the stocks cited in the news in $E(n)_{\Delta t}$. Every position is then closed at the next day's closing price. In our example, on Tuesday at closing time, we would close all the positions opened on Monday. We choose this Δt because our price dataset only includes daily closing prices; consequently, we cannot test Δt shorter than the time between two market closures.

Whenever in a news article or in an idiosyncratic sentence more than one RIC is mentioned, for simplicity we will allocate the same sentiment score to all the RICs mentioned in that news article or part of news. During a period Δt, if a company is mentioned in more than one news article and we have different sentiment scores for the same company, we will count the items in Δt that have a positive (i.e. $S(news) > 0$) and negative sentiment score (i.e. $S(news) < 0$) and long or short the stock, depending on which count is higher. For example, if during a period Δt we have two news articles or idiosyncratic parts of news with a positive sentiment score for the RIC "JPM.N" and one news article or idiosyncratic part of news with a negative sentiment score, we will open a long position for "JPM.N". If the number of positive and negative news articles or idiosyncratic parts of news is equal, we will not open any position for that stock.

4.5 Evaluation Metrics

To evaluate the performance of our investment strategy, we use the following indicators.

Annualised Sharpe ratio (S), defined as:

$$S = \frac{\overline{r}}{\sigma(r)} * \sqrt{D} \tag{4.6}$$

where $\overline{r}$ denotes the mean of all the daily log-returns of our portfolio, $\sigma(r)$ is the standard deviation of all the daily log-returns of our portfolio, and D is the number of trading days in a year[12].

Compound annual growth rate (CAGR), defined as:

$$CAGR = \left(\frac{EV}{BV}\right)^{\frac{1}{n}} - 1 \tag{4.7}$$

[11] As already mentioned, since FinBERT can only classify text with no more than 512 tokens at a time, when calculating $P_+(body)$ we split the body of the news into single sentences, and we calculate the total sentiment score of the news using the equally-weighted average of the positive sentiment score of each sentence in the body. Therefore $P_+(body)$ will be equal to $\frac{1}{N}\sum_{n=1}^{N} P_+(n)$ where $P_+(n)$, is the positive sentiment score of the n^{th} sentence in the body of the news and N is the total number of sentences in the news article.

[12] We choose 252 as the number of trading days for one year.

where EV is the ending value of our portfolio, BV is the beginning value of our portfolio, and n is the number of years between BV and EV.

Maximum drawdown (MDD), defined as:

$$MDD = \max_{\tau \in (0,T)} \left(\max_{t \in (0,\tau)} (P(t) - P(\tau)) \right) \tag{4.8}$$

where MDD at time T indicates the worst loss from peak to trough among successive declines in a time interval $(0, T)$ (J. Choi, 2015). $P(t)$ is the log-value of our portfolio at time t.

Calmar ratio, defined as:

$$Calmar\ ratio = \frac{CAGR}{MDD} \tag{4.9}$$

While the original Calmar Ratio introduced by Young (1991) uses the average annual rate of return for the last 36 months divided by the MDD for the last 36 months, we will use the compound annual growth rate and the maximum drawdown since inception.

5 Results

In this section, we will first provide a complete comparison between the two approaches we use to extract the idiosyncratic part from the news by backtesting them using our news and price data. Next, we will compare the performance of our investment strategy using the idiosyncratic part of news to using the headline or the body of the news. In the end, we will discuss the limitations of our model. No transaction costs are considered in our analysis.

5.1 Comparison between Idiosyncratic Approaches

In Chapter 4, we presented two approaches to extract the idiosyncratic parts from the news text. To compare the two approaches, we test the performance of both approaches on random samples of our dataset and study the sensitivity of both approaches to their input parameters. For the embedding approach, we analyse the sensitivity of our evaluations metrics to the following parameters:

- the measure that we use to calculate the similarity score between our set of clauses and the daily *market component* set of news (i.e. $C^{\text{median}}_{\Delta t}$, $C^{\text{min}}_{\Delta t}$, and $C^{\text{mean}}_{\Delta t}$) (4.2);
- the m^{th} percentile $P_{m\Delta t}$ that we use as similarity threshold between the idiosyncratic clauses and the market component for our daily set of idiosyncratic news $I(m)_{\Delta t}$ (4.3).

For both the embedding and the syntax approach, we test the sensitivity of the evaluation metrics to the n^{th} percentile $P_{n\Delta t}$ that we use as sentiment score threshold to define the set of news we will trade on $E(n)_{\Delta t}$ (4.5). In particular, for our sensitivity analysis, we test 20 subsamples of our dataset, each sample containing news from 365 randomly picked days[1].

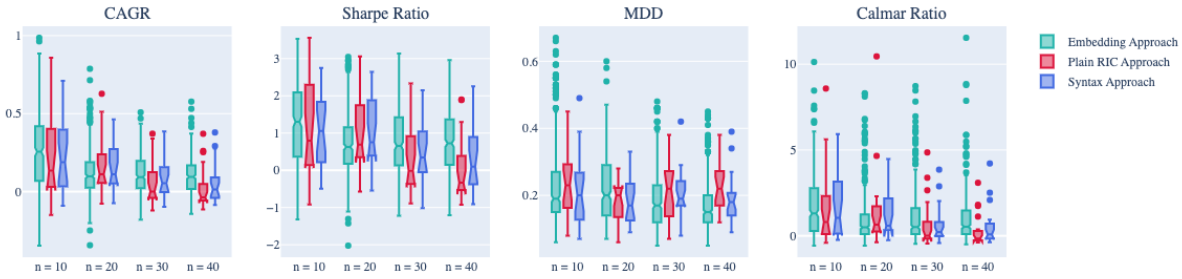

Figure 5.1: Sensitivity of the embedding, syntax, and plain RIC approach to the percentile $P_{n\Delta t}$ that we use as sentiment score threshold for the daily set of news we trade on $E(n)_{\Delta t}$ (4.5). The results shown are based on 20 subsamples of news, each consisting of 365 randomly selected non-consecutive days. For each sub-sample and each n, the embedding approach contains results from 18 scenarios, i.e. for each sub-sample and for each $n \in \{10, 20, 30, 40\}$ we test three different similarity measures ($C^{\text{median}}_{\Delta t}$, $C^{\text{min}}_{\Delta t}$, and $C^{\text{mean}}_{\Delta t}$) (4.2), and for each similarity measure we test six possible percentiles $P_{m\Delta t}$ with $m \in \{1, 5, 10, 20, 30, 40\}$ as similarity thresholds to define the daily set of idiosyncratic news $I(m)_{\Delta t}$ (4.3).

In Figure 5.1, we test the sensitivity of the embedding and syntax approaches to the percentile $P_{n\Delta t}$ that we use to define the daily sentiment score threshold for the set of news we will trade on $E(n)_{\Delta t}$ (4.5). The two approaches are benchmarked to the "plain RIC approach" which consists of using as *idiosyncratic set* of news all the sentences that contain at least a RIC, without applying any additional filter (i.e. we do not check if the company is the subject or the object of the sentence and we do not check the similarity of the clauses to the *market component*). This way, we want to test if the approaches we propose significantly over-perform a "simpler" approach consisting of only selecting the news sentences containing at least a RIC and using it as the idiosyncratic set of

[1] Our dataset of news contains data on 4253 days; therefore, each subsample is roughly 9% of the entire dataset.

news. In addition to the box plot, we run a Kolmogorov-Smirnov (KS) two-sample test for each possible combination of approaches and ns (in Table B.2 we show the complete test results). The two-sample KS test compares the cumulative distributions of two data sets. A two-sample KS statistic for two empirical distribution functions identifies the supremum of the vertical distance set between the two distribution functions (Conover, 1999), i.e. the greatest absolute difference between the two distribution functions. Based on the test's p-values, we can reject or fail to reject the null hypothesis that two samples come from a population with the same distribution. We find that for $n = 30$, when comparing the results distribution functions between the embedding approach and the syntax approach, we can reject the null hypothesis for all the metrics except for the maximum drawdown, while for $n = 40$, the results distribution functions for the embedding approach are significantly different from the ones of the syntax and plain RIC approach for all the metrics except for the maximum drawdown. For the embedding approach, the results of the KS two-samples tests show that using $n = 10$, the evaluation metrics have significantly different distributions compared to using $n \in \{20, 30, 40\}$. For the syntax approach, we see a significant difference in the results' distribution for most of the evaluation metrics when using $n \in \{10, 20\}$ compared to $n = 40$. For the plain RIC approach, we see a significant difference in the results' distribution for most of the evaluation metrics when using $n \in \{30, 40\}$ compared to $n = 10$. In the following analysis and graphs, we will use $n = 10$. In Appendix C we analyse the results using $n \in \{20, 30, 40\}$.

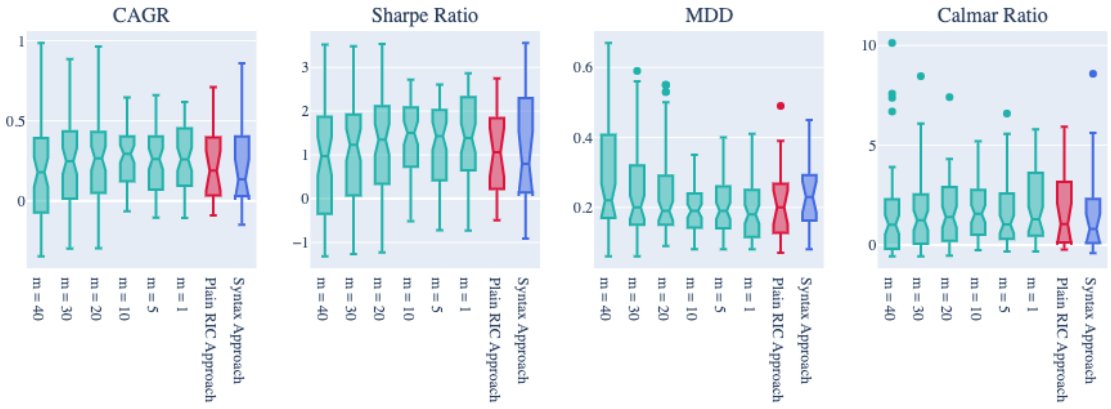

Figure 5.2: Sensitivity of the embedding approach to the percentile m that we use as similarity threshold between idiosyncratic clauses and the market component (4.3). In light-green, we have the results from the embedding approach using different ms, we benchmark them to the syntax, and the plain RIC approaches. The results are based on 20 sub-samples of news, each consisting of 365 randomly selected non-consecutive days. We use $n = 10$ for the sentiment score threshold $P_{n\Delta t}$ (4.5). For the embedding approach, for each of the 20 sub-samples and for each m, we run three scenarios, one for each similarity measures (i.e. $C^{\text{median}}_{\Delta t}$, $C^{\text{min}}_{\Delta t}$, and $C^{\text{mean}}_{\Delta t}$) (4.2).

In Figure 5.2, we show how the performance of the embedding approach changes using different percentiles $P_{m\Delta t}$ as as similarity threshold between idiosyncratic clauses and the market component (4.3) compared to the plain RIC and the syntax approaches. Running some two-samples KS tests, for the embedding approach we find a significant difference in the distribution of the samples only when using $m = 40$ compared to using $m \in \{1, 5, 10\}$. However, we see no significant difference in the distribution of the results between the embedding, syntax and plain RIC approaches for any of the ms. The full results of the tests can be checked in Appendix B, in Table B.3. From now on, we will use $m = 10$ for our similarity threshold $P_{m\Delta t}$ to calculate our daily set of idiosyncratic clauses $I(m)_{\Delta t}$ (4.3).

In Figure 5.3, we show how, for the embedding approach, results change depending on the measure used as similarity score for our set of clauses. Running some KS two-sample tests, we see that the distributions of the evaluation metrics are not significantly different from each other, even compared to the plain RIC approach and the syntax approach. The full results of the tests can be checked in Table B.4. For simplicity, in the next section, for the embedding approach, we will use

$C^{\min}_{\Delta t}$ (4.2) as measure to calculate our daily set of idiosyncratic clauses $I(m)_{\Delta t}$ (4.3) with $m = 10$.

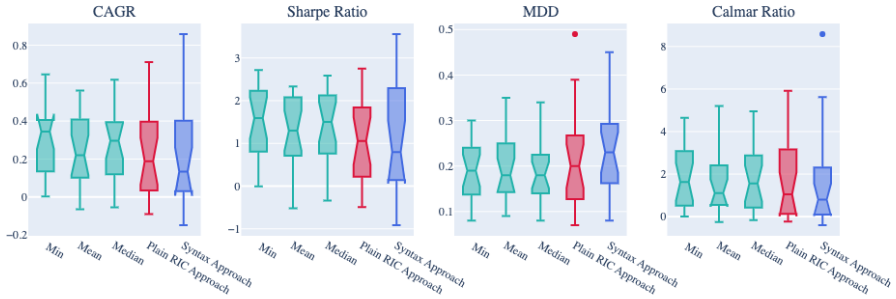

Figure 5.3: Sensitivity of the embedding approach to the measure we use as similarity score for our set of clauses (i.e. $C^{\text{median}}_{\Delta t}$, $C^{\min}_{\Delta t}$, and $C^{\text{mean}}_{\Delta t}$) (4.2). In light green, we have the results from the embedding approach. The results are based on 20 sub-samples of news, each consisting of 365 randomly selected non-consecutive days. We use $n = 10$ for the sentiment score threshold $P_{n\,\Delta t}$ to define the set of news we trade on $E(n)_{\Delta t}$ (4.5) and $m = 10$ for the similarity threshold $P_{m\,\Delta t}$ to calculate our daily set of idiosyncratic clauses $I(m)_{\Delta t}$ (4.3).

5.2 Investment Strategy Results

In Figure 5.4, we show how our sentiment-based investment strategy introduced in Chapter 4 performs when using the idiosyncratic part of news compared to using the headline or the entire news body. As a benchmark, we use a buy-and-hold strategy with the capitalisation-weighted index we created in Chapter 3. We backtest our investment strategy on our set of news between 2007-01-18 and 2018-09-22.

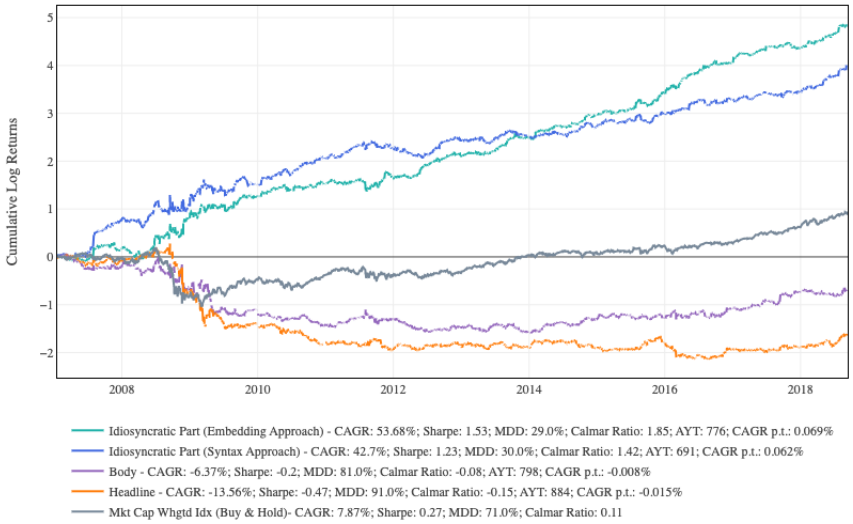

Figure 5.4: Comparison of the cumulative log-returns of our investment strategy using the idiosyncratic parts, the headlines and the bodies of news between 2007-01-18 and 2018-09-22. The grey line represents the benchmark buy-and-hold strategy for the capitalisation-weighted index. The legend shows the average cumulative annual growth rate (CAGR), the annualised Sharpe ratio, the maximum drawdown (MDD), the Calmar ratio, the average yearly roundtrip trades (AYT), and the CAGR per roundtrip trade. We use $n = 10$ to define the sentiment threshold for the daily set of news we trade on $E(n)_{\Delta t}$ (4.5) (meaning that every day we trade only on the stocks which news are within the top 10% most positive/negative news). For the embedding approach we use $C^{\min}_{\Delta t}$ (4.2) and $m = 10$ for our daily set of idiosyncratic news $I(m)_{\Delta t}$ (4.3)

In Figure B.2, we show how the results change for the embedding approach when using different ms and different measures for the similarity score. In Table 5.1, we see how the results change when using a different n to define the sentiment threshold $P_{n\,\Delta t}$ for our daily set of news we trade on $E(n)_{\Delta t}$ (4.5).

Table 5.1: Compound annual growth rate (CAGR), Sharpe ratio, maximum drawdown (MDD), Calmar ratio, average yearly roundtrip trades (AYT), and CAGR per roundtrip trade of our investment strategy between 2007-01-18 and 2018-09-22 by using the idiosyncratic part of news, the headline and the body of news articles. We test different percentiles $P_{n\,\Delta t}$ as sentiment score thresholds (4.5), where Δt is the time between the last market close and twenty minutes before the next market close. Finally, as a benchmark, we present the results of buying and holding our capitalisation-weighted index in the same period.

n	Part of News	CAGR	Sharpe Ratio	MDD	Calmar Ratio	AYT	CAGR p.t.
10	Idiosyncratic Part (Syntax Approach)	42.7%	1.23	30%	1.42	691	0.062%
	Idiosyncratic Part (Embedding Approach)*	53.7%	1.53	29%	1.85	776	0.069%
	Headline	-13.6%	(0.47)	91%	(0.15)	884	-0.015%
	Body	-6.4%	(0.20)	81%	(0.08)	798	-0.008%
20	Idiosyncratic Part (Syntax Approach)	30.2%	1.08	39%	0.77	1,128	0.027%
	Idiosyncratic Part (Embedding Approach)*	34.0%	1.20	33%	1.03	1,281	0.027%
	Headline	-6.2%	(0.23)	75%	(0.08)	1,353	-0.005%
	Body	6.8%	0.20	58%	0.12	1,283	0.005%
30	Idiosyncratic Part (Syntax Approach)	16.3%	0.65	37%	0.44	1,485	0.011%
	Idiosyncratic Part (Embedding Approach)*	28.1%	1.16	29%	0.97	1,680	0.017%
	Headline	-7.0%	(0.28)	80%	(0.09)	1,708	-0.004%
	Body	-3.2%	(0.10)	69%	(0.05)	1,632	-0.002%
40	Idiosyncratic Part (Syntax Approach)	9.5%	0.42	49%	0.19	1,792	0.005%
	Idiosyncratic Part (Embedding Approach)*	20.8%	0.94	28%	0.74	2,006	0.010%
	Headline	-6.0%	(0.24)	77%	(0.08)	2,022	-0.003%
	Body	-1.7%	(0.06)	69%	(0.02)	1,952	-0.001%
	Capitalization-weighted Index	7.9%	0.27	71%	0.11		

* We use $C^{\min}_{\Delta t}$ as similarity measure and $m = 10$ for the daily similarity threshold $P_{m\,\Delta t}$ of the idiosyncratic clauses (4.3).

As we already saw in the previous section, for the idiosyncratic approaches, the choice of the n has a meaningful impact on the results of our investment strategy. The broader the percentile set, the worse are the results for both idiosyncratic approaches. However, for the headline and the body of the news, there does not seem to be a clear relationship between the choice of n and the results of our investment strategy. To confirm that, we run a robustness check on 20 sub-sample, each containing 365 randomly selected days from our set of news for the headline and body approaches. We run some KS two-samples tests to evaluate if using a different n significantly changes the distribution of the investment results of the headline and body approaches. In Figure 5.5, we show the distributions of the results in our random samples. From the KS two-samples test, we find out that there is no significant difference in the distribution of the investment results using different sentiment percentiles $P_{n\,\Delta t}$. In Table B.5 we show the complete results of the KS tests.

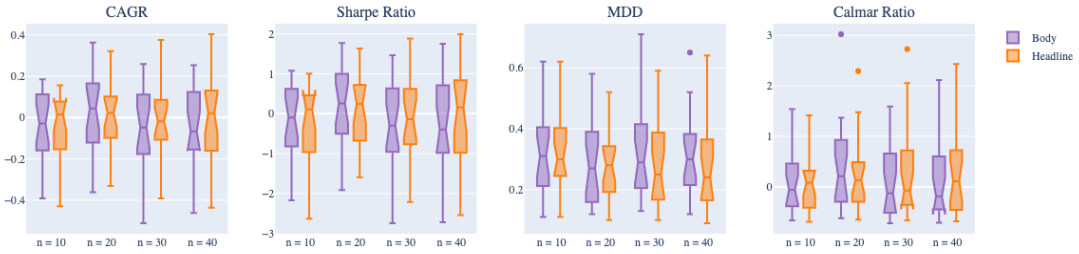

Figure 5.5: Robustness checks for our investment strategies on 20 sub-samples, each of them containing news from 365 randomly selected days, using the headlines and bodies of the news. We test different ns for the sentiment score threshold $P_{n\,\Delta t}$ used to define the set of news we trade on $E(n)_{\Delta t}$ (4.5).

In Table 5.1, we saw that backtesting our investment strategy on the news dataset, overall, the

idiosyncratic approaches beat the headline and body approaches. To confirm our results, we run a robustness check. We create 20 sub-sample, each containing 365 randomly selected days from our set of news, and we compare the performance of the idiosyncratic approaches to the performance of the headline and body approaches. In Figure 5.6, we show the distribution of each of the four approaches for each evaluation metric. We see that using the idiosyncratic part of news as our investment signal generally leads to better results than using the body or the headlines for all the evaluation metrics that we use.

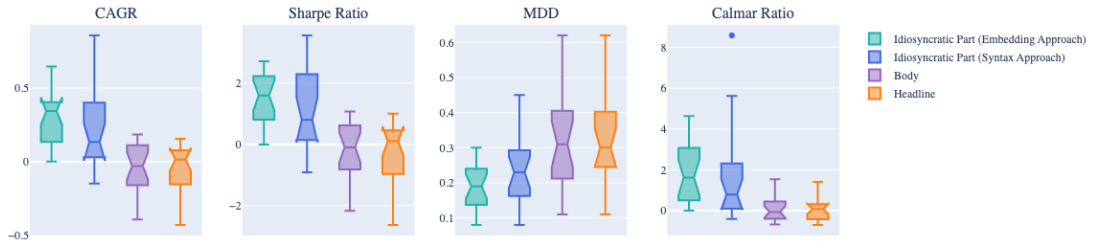

Figure 5.6: Robustness checks for our investment strategies on 20 sub-samples, each of them containing news from 365 randomly selected days. We use $n = 10$ for the sentiment score threshold $P_{n\,\Delta t}$ to define the set of news we trade on $E(n)_{\Delta t}$ (4.5) and for the embedding approach we use $m = 10$ for the similarity threshold $P_{m\,\Delta t}$ to calculate our daily set of idiosyncratic clauses $I(m)_{\Delta t}$ (4.3).

To check if the outperformance of the idiosyncratic approaches is significant, we test if the result samples are drawn from the same continuous distribution of the results from the headline and body approaches performing a two-sample KS test for each measure and each combination of samples. In Table 5.2, we show the statistics and significance levels of the Kolmogorov-Smirnov tests, and we see that in the robustness check, using the idiosyncratic part of the news, the distribution of our evaluation metrics are significantly different compared to using the headline or the body of the news for our investment strategy, meaning that overall using idiosyncratic part of news leads to significantly better results compared to using the headline or the body of the news. However, we do not find a significant difference in the distributions of the results between the two idiosyncratic approaches.

Table 5.2: Statistics and significance levels of the Kolmogorov-Smirnov two-sample tests applied to our robustness check. The null hypothesis is that two samples are taken from populations with identical distributions.

Metric	Part	Body	Headline	Idiosyncratic Part (Embedding Approach)	Idiosyncratic Part (Syntax Approach)
CAGR	Body	0	0.238	0.667***	0.429**
	Headline	0.238	0	0.762***	0.571***
	Idiosyncratic Part (Embedding Approach)	0.667***	0.762***	0	0.286
	Idiosyncratic Part (Syntax Approach)	0.429**	0.571***	0.286	0
Sharpe Ratio	Body	0	0.238	0.667***	0.429**
	Headline	0.238	0	0.762***	0.667***
	Idiosyncratic Part (Embedding Approach)	0.667***	0.762***	0	0.333
	Idiosyncratic Part (Syntax Approach)	0.429**	0.667***	0.333	0
MDD	Body	0	0.143	0.571***	0.381*
	Headline	0.143	0	0.571***	0.381*
	Idiosyncratic Part (Embedding Approach)	0.571***	0.571***	0	0.238
	Idiosyncratic Part (Syntax Approach)	0.381*	0.381*	0.238	0
Calmar Ratio	Body	0	0.143	0.667***	0.476**
	Headline	0.143	0	0.714***	0.524***
	Idiosyncratic Part (Embedding Approach)	0.667***	0.714***	0	0.333
	Idiosyncratic Part (Syntax Approach)	0.476**	0.524***	0.333	0

The significance is reported for the following levels: * p<0.1; ** p< 0.05; *** p<0.01.

5.3 Limitations

Our results showed that using the sentiment score of firm-specific parts of news as a trading signal leads to better performances in our news dataset compared to using the headline or the entire body of the news, supporting our hypothesis. However, we recognise some limitations in our approach that could be addressed in future works.

Besides checking if a piece of information is firm-specific, in our approach we do not analyse the content and the type of information contained in the news. Information like:

> "Shares of Pandora Media Inc (P.N) fell 17 percent to $10.47 following media reports that Apple Inc (AAPL.O) was in talks to license music for a radio service like the one Pandora operates."

even though it is firm-specific and it has a strong negative sentiment, it does not give any information on future price developments; instead, it reflects past stock price movements. In our study, we do not filter out any news based on the type of information provided; therefore, in our approach, information on past prices movements could be used to make predictions on future price movements.

If we take into consideration the following sentence:

> "Research firms say AWS has more than 30 percent of the fast-growing cloud-computing market and it remains far ahead of rivals including Microsoft and Google."

and we calculate the sentiment using FinBERT, we will get a positive sentiment. However, while this certainly is a positive information for Amazon's subsidiary AWS[2], it is not good news for Microsoft[3] and Google[4]. One additional limitation of our study is that whenever different companies are mentioned within the same headline, body or sentence, like in the case above, we assign the same sentiment score to all the companies mentioned. This means that our investment decision, based on the previous news example, would be to invest in both Amazon, Microsoft, and Google stocks equally, even though the positive sentiment should be linked only to Amazon's AWS.

Particularly for our idiosyncratic parts of news extraction methods, we acknowledge one main limitation of both models: as a baseline for both the embedding and the syntax approaches, we use only the sentences in the text that contain a RIC. For example, if a sentence in the body of the article would contain firm-specific information but would not contain any RIC, using our approaches, this sentence would not be processed by our idiosyncratic part of news extraction methods. As explained in Chapter 4, the reason behind it is that in the absence of RICs it is not a trivial task to link firm names to their tickers, given that in our dataset of news, hundreds of different firms are mentioned.

[2] https://aws.amazon.com/
[3] https://www.microsoft.com/
[4] https://www.google.com/

6 Conclusions

In this study, we proposed two methods to disentangle the idiosyncratic part of the news from news articles and showed that using the sentiment score of firm-specific parts of news as a trading signal can lead to better investment results compared to using the sentiment score of the headlines or the bodies of the news to predict next-day stock price movements. The methods we proposed use state-of-the-art NLP techniques to study the syntax and the semantic of the sentences in the news. In the first approach, we analyse the syntactical dependencies of the sentences in the news: only the sentences where a company is either the subject or the object of the sentence are considered idiosyncratic. In the second approach, we discard the parts of news that share a certain degree of semantic similarity with the sentences that represent the market component, and we only retain the idiosyncratic clauses.

To test our trading signal, we proposed a sentiment-based investment strategy that every day takes the news with the strongest sentiment (both positive and negative) and invests in the stocks mentioned in that news. We backtested our investment strategy using a large set of news articles from Reuters published between 2007 and 2018. We compared the investment results using the sentiment scores of the idiosyncratic parts, the headlines and the news bodies. We found that using the idiosyncratic part of news (either of the two approaches we introduced) yields significantly better results than using either the headline or the whole news body when we backtest these on our set of news.

As a benchmark, we compared the two idiosyncratic approaches to the plain RIC approach to determine if one approach systematically outperforms the other. When we backtested the two approaches on random news samples, we did not find significant differences in the results' distributions between the two approaches we proposed nor with the benchmark, meaning that neither approach is systematically superior or inferior to the other or to the benchmark.

We showed that, for the idiosyncratic approaches, the choice of the percentile $P_{n\Delta t}$ that we use as a sentiment score threshold to define the set of news we trade on $E(n)_{\Delta t}$ (4.5) plays a meaningful role in the performance of our investment strategy. The smaller the n, the fewer the number of news we trade on each day, the stronger the sentiment signal, the better the performance of our investment strategy. On the contrary, the choice of n does not significantly change the results when using the news headlines and bodies for our investment strategy.

In the end, we highlighted some limitations of our model, which can be addressed in future work.

References

Alammar, J. (2018, Dec). *The illustrated bert, elmo, and co. (how nlp cracked transfer learning).* Retrieved from https://jalammar.github.io/illustrated-bert/

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Chan, S. W., & Franklin, J. (2011). A text-based decision support system for financial sequence prediction. *Decision Support Systems*, *52*(1), 189–198.

Chen, D., Zou, Y., Harimoto, K., Bao, R., Ren, X., & Sun, X. (2019). Incorporating fine-grained events in stock movement prediction. *arXiv preprint arXiv:1910.05078*.

Chen, Q. (2021). Stock movement prediction with financial news using contextualized embedding from bert. *arXiv preprint arXiv:2107.08721*.

Choi, J. (2015). Maximum drawdown, recovery and momentum. *Recovery and Momentum (March 31, 2015)*.

Choi, J. D., Chen, H., & Jurczyk, T. (2016, Jan). *Clearnlp dependency labels.* Retrieved from https://github.com/clir/clearnlp-guidelines/blob/master/md/components/dependency_parsing.md

Conover, W. J. (1999). *Practical nonparametric statistics* (Vol. 350). john wiley & sons.

Dai, A. M., & Le, Q. V. (2015). *Semi-supervised sequence learning.*

Daniel, G., Sornette, D., & Woehrmann, P. (2009). Look-ahead benchmark bias in portfolio performance evaluation. *The Journal of Portfolio Management*, *36*(1), 121–130. Retrieved from https://jpm.pm-research.com/content/36/1/121 doi: 10.3905/JPM.2009.36.1.121

DeLisle, R. J., Mauck, N., & Smedema, A. R. (2016). Idiosyncratic volatility and firm-specific news: Beyond limited arbitrage. *Financial Management*, *45*(4), 923–951.

Deng, S., Zhang, N., Zhang, W., Chen, J., Pan, J. Z., & Chen, H. (2019). Knowledge-driven stock trend prediction and explanation via temporal convolutional network. In *Companion proceedings of the 2019 world wide web conference* (pp. 678–685).

Devlin, J., & Chang, M.-W. (2018, Nov). *Open sourcing bert: State-of-the-art pre-training for natural language processing.* Retrieved from https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2014). Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1415–1425).

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence.*

Engle, R. F., Hansen, M. K., Karagozoglu, A. K., & Lunde, A. (2021). News and idiosyncratic volatility: The public information processing hypothesis. *Journal of Financial Econometrics*, *19*(1), 1–38.

Fama, E. F. (1965). The behavior of stock-market prices. *The journal of Business*, *38*(1), 34–105.

Genc, Z., & Araci, D. T. (2020, Jul). *Finbert: financial sentiment analysis with bert.* Prosus. Retrieved from https://www.prosus.com/news/finbert-financial-sentiment-analysis-with-bert/

Howard, J., & Ruder, S. (2018). *Universal language model fine-tuning for text classification.*

Hui, J. L. O., Hoon, G. K., & Zainon, W. M. N. W. (2017). Effects of word class and text position in sentiment-based news classification. *Procedia Computer Science*, *124*, 77–85.

Ke, Z. T., Kelly, B. T., & Xiu, D. (2019, August). *Predicting returns with text data* (Working Paper No. 26186). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w26186 doi: 10.3386/w26186

Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of hong kong. *Information Processing & Management*, *57*(5), 102212.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality.*

Narayan, P. K., & Bannigidadmath, D. (2017). Does financial news predict stock returns? new evidence from islamic and non-islamic stocks. *Pacific-Basin Finance Journal*, *42*, 24–45.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations.*

Qiao, Y., Xiong, C., Liu, Z., & Liu, Z. (2019). Understanding the behaviors of bert in ranking. *arXiv preprint arXiv:1904.07531*.

Reimers, N., & Gurevych, I. (2019). *Sentence-bert: Sentence embeddings using siamese bert-networks.*

Ryan, P., & Taffler, R. J. (2004). Are economically significant stock returns and trading volumes driven by firm-specific news releases? *Journal of Business Finance & Accounting*, *31*(1-2), 49–82.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, *27*(2), 1–19.

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020, September). Mpnet: Masked and permuted pre-training for language understanding. In *Neurips 2020.* Retrieved from https://www.microsoft.com/en-us/research/publication/mpnet-masked-and-permuted-pre-training-for-language-understanding/

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, *62*(3), 1139–1168.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). *Attention is all you need.*

Virlics, A. (2013). Investment decision making and risk. *Procedia Economics and Finance*, *6*, 169–177.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... others (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, *abs/1609.08144*. Retrieved from http://arxiv.org/abs/1609.08144

Yang, Y., Uy, M. C. S., & Huang, A. (2020). Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Young, T. W. (1991). Calmar ratio: A smoother tool. *Futures*, *20*(1), 40.

A BERT and FinBERT

The year 2018 represents a milestone for natural language processing (NLP) thanks to the release of the open-source technique for NLP pre-training called bidirectional encoder representations from transformers (BERT). While BERT builds upon precedent works in pre-training contextual representations, for example, semi-supervised sequence learning (Dai & Le, 2015), generative pre-training[1], ELMo (Peters et al., 2018), and ULMFit (Howard & Ruder, 2018), BERT is the first *deeply bidirectional unsupervised* language model, pre-trained on a plain text corpus[2] to predict masked words.

NLP models can be pre-trained either in a *context-free* or *contextual* way, and *contextual* representations can be classified into *unidirectional* or *bidirectional*. Context-free representations such as word2vec (Mikolov, Sutskever, et al., 2013; Mikolov, Chen, et al., 2013) and Glove (Pennington et al., 2014) represent each word in a vocabulary with a single word embedding (Devlin & Chang, 2018). For example, in context-free representation, the word "prices" would have the same representation in "import prices rise in February on higher oil" and in "trading was choppy but prices clawed back briefly in the early afternoon". Contextual models instead generate a representation of each word based on the other words contained in the sentence, for example, in the sentence "Trading was choppy but prices clawed back briefly in the early afternoon", a unidirectional contextual model would represent "prices" based on "Trading was choppy but" and not "clawed back briefly in the early afternoon". Since BERT uses a contextual *bidirectional* representation of words, it will represent "prices" using both its previous and following context (i.e. "Trading was choppy but ... clawed back briefly in the early afternoon"). In Figure A.1, we visualise BERT's neural network architecture compared to generative pre-training (OpenAI GPT) and ELMo.

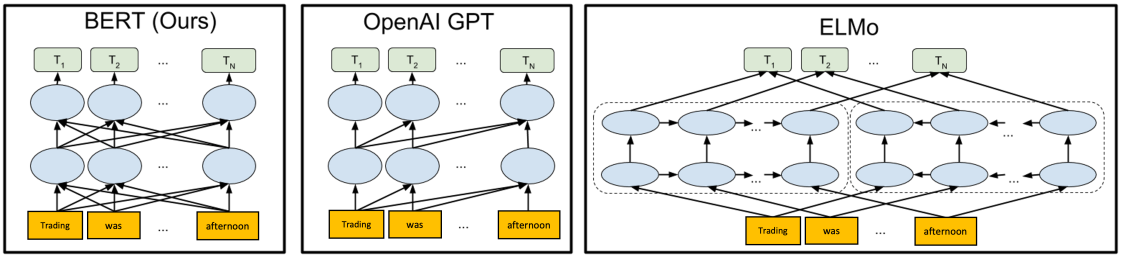

Figure A.1: Adapted from (Devlin & Chang, 2018). Comparison between BERT (bidirectional), OpenAI GPT (unidirectional) and ELMo (shallowly bidirectional). The arrows indicate how the information flows from one layer to the other, the green boxes at the top are the contextualized representation of each input word.

To efficiently train a bidirectional model, masked language modelling (MLM) and next sentence prediction (NSP) have been used. The first method, the MLM, consists of randomly masking a certain percentage of words in a sentence by replacing them with the token "[MASK]" and, by conditioning each word in the sentence bidirectionally, the masked word is predicted. For example, giving as an input to the model the sentence "Crude [MASK] prices have gained about 9 per cent in the following month.", the model will predict that "[MASK]" = "oil". The second method, the NSP, is used to understand the relationship between sentences. The model is trained to predict if the second sentence is the following sentence to the first one in a pair of sentences provided as inputs.

BERT models are usually pre-trained on a large corpus of text, then fine-tuned for specific tasks. One way to use BERT is to classify pieces of text into classes, such as sentiment labels (e.g. positive, negative and neutral). To fine-tune such models, we would need to train the classifier (i.e. a feed-forward neural network with a final softmax layer) on a labelled dataset. In the case

[1] https://openai.com/blog/language-unsupervised/
[2] BERT has been pre-trained on Wikipedia's corpus.

of sentiment analysis, a labelled dataset would be a list of sentences with a label for each sentence indicating the sentiment (e.g. "positive", "negative" or "neutral").

Devlin et al. (2018) present two model sizes for BERT: $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$. The difference between the two models stands on the number of encoder layers (or "transformer blocks"): for $\text{BERT}_{\text{BASE}}$ we have 12 layers, while for $\text{BERT}_{\text{LARGE}}$ we have 24 layers.

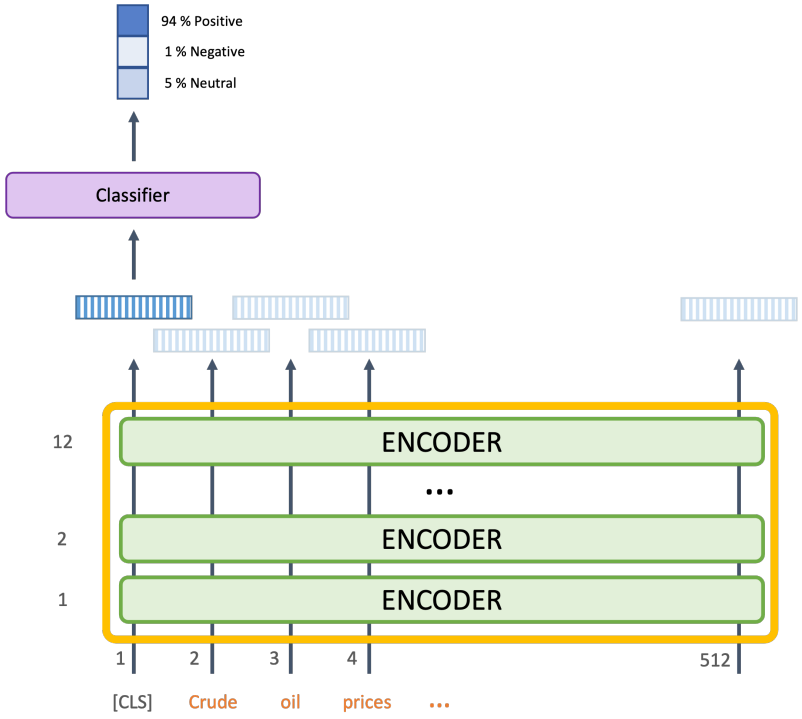

Figure A.2: Illustration of how $\text{BERT}_{\text{BASE}}$ works in a sentiment classification task.

In Figure A.2, we show how the BERT model works. In the case of a sentiment classification problem (e.g. our study), the model's input is a sequence of tokens, starting from a "[CLS]" token followed by a token for each word. Following the normal functioning of a Transformer[3], for each encoder, the input passes through a self-attention layer and its output is fed into a feed-forward neural network which will hand in the output to the next encoder in the stack and at the end of the stack. In the end, the stack of encoders will output a vector of size *hidden_size* (768 for $\text{BERT}_{\text{BASE}}$). The last hidden state of the special "[CLS]" token is used for the classifications task. This is used as input for the fine-tuned classifier, made by a feed-forward neural network and a final softmax layer that gives the probability of class membership for each label (Alammar, 2018).

FinBERT is a BERT-based language model with a better understanding of financial language that has been fine-tuned for sentiment classification. Araci (2019) further pre-trained BERT on a purely financial corpus from Reuters TRC2[4] in order to adapt the general domain of BERT to the financial domain and fine-tuned with labeled data for financial sentiment classification (Genc & Araci, 2020).

[3] For an extensive explanation on how transformers work, please refer to the work of Vaswani et al. (2017).
[4] https://trec.nist.gov/data/reuters/reuters.html

B Additional Results

B.1 Results with Neutral News

In Figure B.1, we show the same results as in Figure 5.4, compared to the results of our investment strategy when not filtering out the news with a neutral sentiment. The results are worse for all the news parts except for the news headline when not filtering out the news with a neutral sentiment. In Table B.1, we show the results of our investment strategy using different ns without filtering out the news with a neutral sentiment.

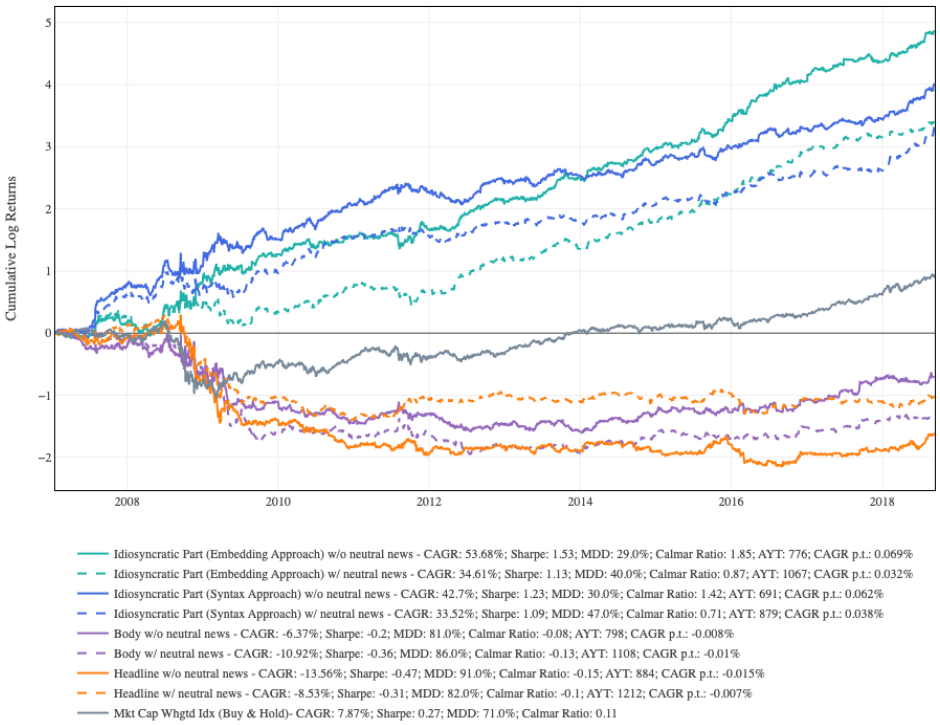

Figure B.1: Comparison of the cumulative log-returns of our investment strategy using the idiosyncratic parts, the headlines and the bodies of news between 2007-01-18 and 2018-09-22. The grey line represents the benchmark buy-and-hold strategy for the capitalisation-weighted index. The legend shows the average cumulative annual growth rate (CAGR), the annualised Sharpe ratio, the maximum drawdown (MDD), the Calmar ratio, the average yearly roundtrip trades (AYT), and the CAGR per roundtrip trade. The dashed lines represent the results when not filtering out the news with neutral sentiments. We use $n = 10$ to define the sentiment threshold for our daily set of news we trade on $E(n)_{\Delta t}$ (4.5) (meaning that every day we trade only on the stocks which news are within the top 10% most positive/negative news). For the embedding approach, we use $C^{\min}_{\Delta t}$ (4.2) and $m = 10$ for our daily set of idiosyncratic news $I(m)_{\Delta t}$ (4.3)

Table B.1: Compound annual growth rate (CAGR), Sharpe ratio and maximum drawdown (MDD), Calmar ratio, average yearly roundtrip trades (AYT), and CAGR per roundtrip trade of our investment strategy between 2007-01-18 and 2018-09-22 by using the idiosyncratic part of the news, the headline and the body of news articles. We compare the results when we filter out the news with a neutral sentiment with those when we do not. We test different percentiles $P_{n\,\Delta t}$ to define the set of news $E(n)_{\Delta t}$ (4.5) on which we will trade, where Δt is the time between the last market close and twenty minutes before the next market close. Finally, as a benchmark, we present the results of buying and holding our capitalisation-weighted index in the same period.

		CAGR		Sharpe Ratio		MDD		Calmar Ratio		AYT		CAGR p.t.	
n	Part of News	w/o neutr	w/ neutr	w/o neutr	w/ neutr	w/o neutr	w/ neutr	w/o neutr	w/ neutr	w/o neutr	w/ neutr	w/o neutr	w/ neutr
10	Idiosyncratic Part (Syntax Approach)	42.7%	33.5%	1.23	1.09	30%	47%	1.42	0.71	691	879	0.062%	0.038%
	Idiosyncratic Part (Embedding Approach)*	53.7%	34.6%	1.53	1.13	29%	40%	1.85	0.87	776	1,067	0.069%	0.032%
	Headline	-13.6%	-8.5%	(0.47)	(0.31)	91%	82%	(0.15)	(0.10)	884	1,212	-0.015%	-0.007%
	Body	-6.4%	-10.9%	(0.20)	(0.36)	81%	86%	(0.08)	(0.13)	798	1,108	-0.008%	-0.010%
20	Idiosyncratic Part (Syntax Approach)	30.2%	19.0%	1.08	0.76	39%	33%	0.77	0.58	1,128	1,451	0.027%	0.013%
	Idiosyncratic Part (Embedding Approach)*	34.0%	23.2%	1.20	1.01	33%	30%	1.03	0.77	1,281	1,729	0.027%	0.013%
	Headline	-6.2%	-10.4%	(0.23)	(0.44)	75%	82%	(0.08)	(0.13)	1,353	1,857	-0.005%	-0.006%
	Body	6.8%	-3.1%	0.20	(0.10)	58%	66%	0.12	(0.05)	1,283	1,755	0.005%	-0.002%
30	Idiosyncratic Part (Syntax Approach)	16.3%	12.7%	0.65	0.58	37%	38%	0.44	0.33	1,485	1,884	0.011%	0.007%
	Idiosyncratic Part (Embedding Approach)*	28.1%	15.4%	1.16	0.71	29%	36%	0.97	0.43	1,680	2,199	0.017%	0.007%
	Headline	-7.0%	-7.9%	(0.28)	(0.34)	80%	78%	(0.09)	(0.10)	1,708	2,305	-0.004%	-0.003%
	Body	-3.2%	-1.8%	(0.10)	(0.06)	69%	67%	(0.05)	(0.03)	1,632	2,211	-0.002%	-0.001%
40	Idiosyncratic Part (Syntax Approach)	9.5%	2.4%	0.42	0.11	49%	51%	0.19	0.05	1,792	2,257	0.005%	0.001%
	Idiosyncratic Part (Embedding Approach)*	20.8%	13.6%	0.94	0.62	28%	29%	0.74	0.47	2,006	2,536	0.010%	0.005%
	Headline	-6.0%	-7.8%	(0.24)	(0.33)	77%	72%	(0.08)	(0.11)	2,022	2,652	-0.003%	-0.003%
	Body	-1.7%	-5.4%	(0.06)	(0.19)	69%	72%	(0.02)	(0.08)	1,952	2,590	-0.001%	-0.002%
	Capitalization-weighted Index	7.9%		0.27		71%		0.11					

* We use $C^{\min}_{\Delta t}$ as similarity measure and $m = 10$ for the daily similarity threshold $P_{m\,\Delta t}$ of the idiosyncratic clauses (4.3).

In Figure B.2, we show how the results change for the embedding approach when using different ms and different measures for the similarity score.

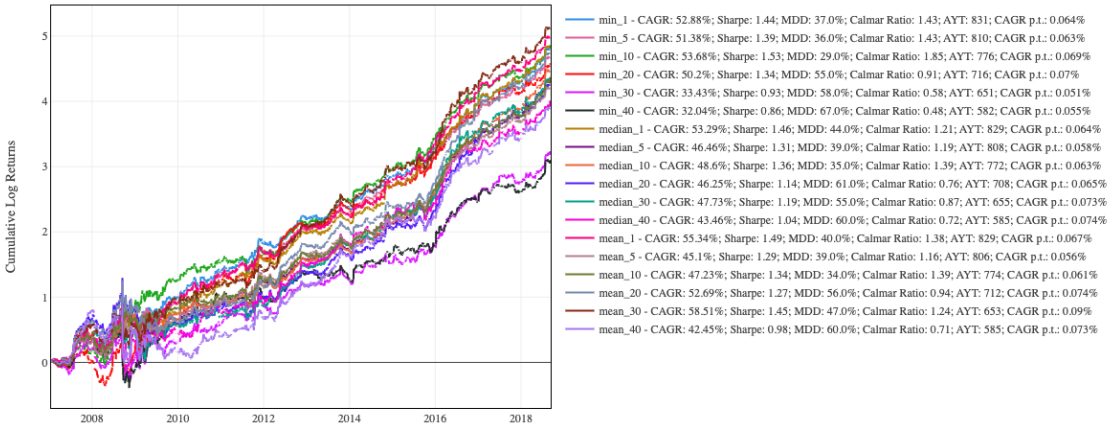

Figure B.2: Comparison of the cumulative log-returns of our investment strategy using different setups of the embedding approach between 2007-01-18 and 2018-09-22. The legend shows the average cumulative annual growth rate (CAGR), the annualised Sharpe ratio, the maximum drawdown (MDD), the Calmar ratio, the average yearly roundtrip trades (AYT), and the CAGR per roundtrip trade. We test three measures of similarity: $C^{\text{median}}_{\Delta t}$, $C^{\min}_{\Delta t}$, and $C^{\text{mean}}_{\Delta t}$ (4.2), respectively called median, min, and mean in the legend of the graph. For each measure, we test six different values of m, used to define the similarity threshold $P_{m\,\Delta t}$ for our daily set of idiosyncratic news $I(m)_{\Delta t}$ (4.3). For example, "mean_1" indicates that we use the measure $C^{\text{mean}}_{\Delta t}$ and $m = 1$.

B.2 Kolmogorov-Smirnov two-sample Tests Results

In this section we present the Kolmogorov-Smirnov two-sample tests results mentioned in Chapter 5.

Table B.2: Kolmogorov-Smirnov two-sample test statistics and significance levels for the percentiles $P_{n\,\Delta t}$ we use as threshold to define the set of news we will trade on (4.5). The results shown are based on 20 sub-samples of news, each consisting of 365 randomly selected non-consecutive days. For each sub-sample and each n, the embedding approach includes 18 scenarios, i.e. for each sub-sample and for each $n \in \{10, 20, 30, 40\}$ we test three different similarity measures ($C^{\text{median}}_{\Delta t}$, $C^{\min}_{\Delta t}$, and $C^{\text{mean}}_{\Delta t}$) (4.2), and for each similarity measure we test six possible percentiles m (i.e. 1, 5, 10, 20, 30, 40) as thresholds to define our daily set of idiosyncratic news (4.3).

Metric	n	Approach	10			20			30			40		
			Embedding Approach	Plain RIC Approach	Syntax Approach	Embedding Approach	Plain RIC Approach	Syntax Approach	Embedding Approach	Plain RIC Approach	Syntax Approach	Embedding Approach	Plain RIC Approach	Syntax Approach
CAGR	10	Embedding Approach	0	0.164	0.206	0.357**	0.265	0.299**	0.362***	0.437***	0.476***	0.413***	0.513***	0.569***
		Plain RIC Approach	0.164	0	0.19	0.317**	0.238	0.238	0.307**	0.429**	0.476**	0.325**	0.524***	0.571***
		Syntax Approach	0.206	0.19	0	0.262	0.19	0.19	0.27*	0.381*	0.429**	0.331**	0.476**	0.524***
	20	Embedding Approach	0.357***	0.317**	0.262	0	0.18	0.177	0.05	0.259	0.373***	0.082	0.426***	0.519***
		Plain RIC Approach	0.265	0.238	0.19	0.18	0	0.143	0.201	0.333	0.429**	0.246	0.524***	0.571***
		Syntax Approach	0.299**	0.238	0.19	0.177	0.143	0	0.177	0.381*	0.476**	0.22	0.476**	0.571***
	30	Embedding Approach	0.362***	0.307**	0.27*	0.05	0.201	0.177	0	0.257	0.373***	0.077	0.415***	0.505***
		Plain RIC Approach	0.437***	0.429**	0.381*	0.259	0.333	0.381*	0.257	0	0.238	0.206	0.286	0.476**
		Syntax Approach	0.476***	0.476**	0.429**	0.373***	0.429**	0.476**	0.373***	0.238	0	0.357***	0.095	0.333
	40	Embedding Approach	0.413***	0.325**	0.331**	0.082	0.246	0.22	0.077	0.206	0.357***	0	0.384***	0.524***
		Plain RIC Approach	0.513***	0.524***	0.476**	0.426***	0.524***	0.476**	0.415***	0.286	0.095	0.384***	0	0.333
		Syntax Approach	0.569***	0.571***	0.524***	0.519***	0.571***	0.571***	0.505***	0.476**	0.333	0.524***	0.333	0
Sharpe Ratio	10	Embedding Approach	0	0.111	0.204	0.304***	0.204	0.233	0.243***	0.402***	0.442***	0.246***	0.46***	0.526***
		Plain RIC Approach	0.111	0	0.19	0.286*	0.238	0.238	0.257	0.429**	0.429**	0.241	0.429**	0.571***
		Syntax Approach	0.204	0.19	0	0.23	0.19	0.238	0.204	0.381*	0.429**	0.235	0.429**	0.476**
	20	Embedding Approach	0.304***	0.286*	0.23	0	0.225	0.164	0.082	0.241	0.378***	0.082	0.394***	0.537***
		Plain RIC Approach	0.204	0.238	0.19	0.225	0	0.143	0.172	0.381*	0.476**	0.193	0.476**	0.571***
		Syntax Approach	0.233	0.238	0.238	0.164	0.143	0	0.127	0.333	0.429**	0.175	0.476**	0.524***
	30	Embedding Approach	0.243***	0.257	0.204	0.082	0.172	0.127	0	0.249	0.368***	0.058	0.399***	0.497***
		Plain RIC Approach	0.402***	0.429**	0.381*	0.241	0.381*	0.333	0.249	0	0.238	0.233	0.286	0.476**
		Syntax Approach	0.442***	0.429**	0.429**	0.378***	0.476**	0.429**	0.368***	0.238	0	0.378***	0.095	0.333
	40	Embedding Approach	0.246***	0.241	0.235	0.082	0.193	0.175	0.058	0.233	0.378***	0	0.373***	0.516***
		Plain RIC Approach	0.46***	0.429**	0.429**	0.394***	0.476**	0.476**	0.399***	0.286	0.095	0.373***	0	0.333
		Syntax Approach	0.526***	0.571***	0.476**	0.537***	0.571***	0.524***	0.497***	0.476**	0.333	0.516***	0.333	0
MDD	10	Embedding Approach	0	0.164	0.127	0.063	0.185	0.243	0.148***	0.177	0.159	0.238***	0.209	0.172
		Plain RIC Approach	0.164	0	0.238	0.164	0.143	0.238	0.153	0.19	0.19	0.249	0.238	0.333
		Syntax Approach	0.127	0.238	0	0.114	0.286	0.286	0.257	0.19	0.143	0.347**	0.286	0.143
	20	Embedding Approach	0.063	0.164	0.114	0	0.196	0.251	0.169***	0.172	0.127	0.257***	0.254	0.169
		Plain RIC Approach	0.185	0.143	0.286	0.196	0	0.19	0.111	0.333	0.238	0.153	0.143	0.286
		Syntax Approach	0.243	0.238	0.286	0.251	0.19	0	0.23	0.143	0.286	0.344**	0.286	0.286
	30	Embedding Approach	0.148***	0.153	0.257	0.169***	0.111	0.23	0	0.28*	0.241	0.114**	0.198	0.302**
		Plain RIC Approach	0.177	0.19	0.19	0.172	0.333	0.143	0.28*	0	0.19	0.386***	0.238	0.19
		Syntax Approach	0.159	0.19	0.143	0.127	0.238	0.286	0.241	0.19	0	0.328**	0.333	0.19
	40	Embedding Approach	0.238***	0.249	0.347**	0.257***	0.153	0.344**	0.114**	0.386***	0.328**	0	0.233	0.392***
		Plain RIC Approach	0.209	0.238	0.286	0.254	0.143	0.286	0.198	0.238	0.333	0.233	0	0.333
		Syntax Approach	0.172	0.333	0.143	0.169	0.286	0.286	0.302**	0.19	0.19	0.392***	0.333	0
Calmar Ratio	10	Embedding Approach	0	0.119	0.225	0.296***	0.233	0.241	0.243***	0.418***	0.442***	0.238***	0.471***	0.524***
		Plain RIC Approach	0.119	0	0.19	0.296**	0.286	0.238	0.278*	0.429**	0.476**	0.251	0.429**	0.571***
		Syntax Approach	0.225	0.19	0	0.18	0.143	0.143	0.138	0.381*	0.381*	0.18	0.429**	0.476**
	20	Embedding Approach	0.296***	0.296**	0.18	0	0.198	0.188	0.085	0.257	0.373***	0.074	0.384***	0.532***
		Plain RIC Approach	0.233	0.286	0.143	0.198	0	0.143	0.183	0.429**	0.476**	0.206	0.524***	0.571***
		Syntax Approach	0.241	0.238	0.143	0.188	0.143	0	0.167	0.381*	0.429**	0.193	0.476**	0.524***
	30	Embedding Approach	0.243***	0.278*	0.138	0.085	0.183	0.167	0	0.28*	0.373***	0.061	0.386***	0.497***
		Plain RIC Approach	0.418***	0.429**	0.381*	0.257	0.429**	0.381*	0.28*	0	0.238	0.243	0.238	0.429**
		Syntax Approach	0.442***	0.476**	0.381*	0.373***	0.476**	0.429**	0.373***	0.238	0	0.368***	0.143	0.381*
	40	Embedding Approach	0.238***	0.251	0.18	0.074	0.206	0.193	0.061	0.243	0.368***	0	0.376***	0.524***
		Plain RIC Approach	0.471***	0.429**	0.429**	0.384***	0.524***	0.476**	0.386***	0.238	0.143	0.376***	0	0.333
		Syntax Approach	0.524***	0.571***	0.476**	0.532***	0.571***	0.524***	0.497***	0.429**	0.381*	0.524***	0.333	0

The significance is reported for the following levels: * p<0.1; ** p< 0.05; *** p<0.01.

Table B.3: Kolmogorov-Smirnov two-sample test statistics and significance levels for the percentiles $P_{m\,\Delta t}$ we use as similarity score for our set of clauses (4.3) in the embedding approach. The results shown are based on 20 sub-samples of news, each consisting of 365 randomly selected non-consecutive days. We use the $n = 10$ as sentiment score threshold for the set of news we trade on $E(n)_{\Delta t}$ (4.5). For the embedding approach, for each of the 20 sub-samples and for each m, we run three scenarios, one for each similarity measures (i.e. $C^{\text{median}}_{\Delta t}$, $C^{\text{min}}_{\Delta t}$, and $C^{\text{mean}}_{\Delta t}$) (4.2).

Metric	m	1	5	10	20	30	40	Syntax Approach	Plain RIC Approach
CAGR	**1**	0	0.127	0.127	0.175	0.238*	0.286**	0.222	0.19
	5	0.127	0	0.143	0.143	0.206	0.286**	0.19	0.143
	10	0.127	0.143	0	0.143	0.175	0.286**	0.254	0.238
	20	0.175	0.143	0.143	0	0.079	0.159	0.254	0.19
	30	0.238*	0.206	0.175	0.079	0	0.175	0.222	0.143
	40	0.286**	0.286**	0.286**	0.159	0.175	0	0.159	0.27
	Syntax Approach	0.222	0.19	0.254	0.254	0.222	0.159	0	0.19
	Plain RIC Approach	0.19	0.143	0.238	0.19	0.143	0.27	0.19	0
Sharpe	**1**	0	0.175	0.159	0.175	0.238*	0.286**	0.254	0.19
	5	0.175	0	0.159	0.143	0.206	0.27**	0.222	0.143
	10	0.159	0.159	0	0.143	0.175	0.27**	0.27	0.159
	20	0.175	0.143	0.143	0	0.095	0.19	0.238	0.159
	30	0.238*	0.206	0.175	0.095	0	0.159	0.222	0.127
	40	0.286**	0.27**	0.27**	0.19	0.159	0	0.159	0.254
	Syntax Approach	0.254	0.222	0.27	0.238	0.222	0.159	0	0.19
	Plain RIC Approach	0.19	0.143	0.159	0.159	0.127	0.254	0.19	0
MDD	**1**	0	0.079	0.143	0.175	0.19	0.254**	0.19	0.159
	5	0.079	0	0.159	0.143	0.159	0.254**	0.159	0.143
	10	0.143	0.159	0	0.175	0.222*	0.302***	0.222	0.143
	20	0.175	0.143	0.175	0	0.111	0.159	0.159	0.159
	30	0.19	0.159	0.222*	0.111	0	0.143	0.127	0.206
	40	0.254**	0.254**	0.302***	0.159	0.143	0	0.222	0.27
	Syntax Approach	0.19	0.159	0.222	0.159	0.127	0.222	0	0.238
	Plain RIC Approach	0.159	0.143	0.143	0.159	0.206	0.27	0.238	0
Calmar	**1**	0	0.143	0.143	0.175	0.206	0.286**	0.222	0.175
	5	0.143	0	0.175	0.143	0.206	0.27**	0.206	0.159
	10	0.143	0.175	0	0.143	0.175	0.254**	0.286	0.19
	20	0.175	0.143	0.143	0	0.095	0.175	0.286	0.159
	30	0.206	0.206	0.175	0.095	0	0.159	0.19	0.143
	40	0.286**	0.27**	0.254**	0.175	0.159	0	0.159	0.254
	Syntax Approach	0.222	0.206	0.286	0.286	0.19	0.159	0	0.19
	Plain RIC Approach	0.175	0.159	0.19	0.159	0.143	0.254	0.19	0

The significance is reported for the following levels: * p<0.1; ** p< 0.05; *** p<0.01.

Table B.4: Kolmogorov-Smirnov two-sample test statistics and significance levels for the measures we use as similarity score for our set of clauses (i.e. min, median, mean) (4.2) in the embedding approach. The results shown are based on 20 sub-samples of news, each consisting of 365 randomly selected non-consecutive days. We use the $n = 10$ as sentiment score threshold for the set of news we trade on $E(n)_{\Delta t}$ (4.5) and $m = 10$ as similarity threshold $P_{m\,\Delta t}$ to calculate our daily set of idiosyncratic clauses $I(m)_{\Delta t}$ (4.3).

Metric	Measure	Min	Mean	Median	Plain RIC Approach	Syntax Approach
CAGR	**Min**	0	0.19	0.19	0.286	0.286
	Mean	0.19	0	0.143	0.19	0.238
	Median	0.19	0.143	0	0.238	0.238
	Plain RIC Approach	0.286	0.19	0.238	0	0.19
	Syntax Approach	0.286	0.238	0.238	0.19	0
Sharpe	**Min**	0	0.238	0.143	0.19	0.333
	Mean	0.238	0	0.19	0.143	0.286
	Median	0.143	0.19	0	0.19	0.286
	Plain RIC Approach	0.19	0.143	0.19	0	0.19
	Syntax Approach	0.333	0.286	0.286	0.19	0
MDD	**Min**	0	0.095	0.143	0.143	0.238
	Mean	0.095	0	0.143	0.143	0.19
	Median	0.143	0.143	0	0.143	0.286
	Plain RIC Approach	0.143	0.143	0.143	0	0.238
	Syntax Approach	0.238	0.19	0.286	0.238	0
Calmar	**Min**	0	0.19	0.143	0.238	0.333
	Mean	0.19	0	0.19	0.143	0.19
	Median	0.143	0.19	0	0.238	0.333
	Plain RIC Approach	0.238	0.143	0.238	0	0.19
	Syntax Approach	0.333	0.19	0.333	0.19	0

The significance is reported for the following levels: * p<0.1; ** p< 0.05; *** p<0.01.

Table B.5: Kolmogorov-Smirnov two-sample test statistics and significance levels for the percentiles $P_{n\,\Delta t}$ we use as sentiment threshold to define the set of news we will trade on (4.5) using the sentiment of the headline and the body of the news for our investment strategy. The results shown are based on 20 sub-samples of news, each consisting of 365 randomly selected non-consecutive days.

			10		20		30		40	
Metric	**n**	**News Part**	**Body**	**Headline**	**Body**	**Headline**	**Body**	**Headline**	**Body**	**Headline**
CAGR	**10**	**Body**	0	0.238	0.238	0.143	0.19	0.19	0.238	0.19
		Headline	0.238	0	0.333	0.19	0.238	0.19	0.238	0.238
	20	**Body**	0.238	0.333	0	0.19	0.286	0.286	0.333	0.19
		Headline	0.143	0.19	0.19	0	0.238	0.143	0.238	0.143
	30	**Body**	0.19	0.238	0.286	0.238	0	0.143	0.143	0.19
		Headline	0.19	0.19	0.286	0.143	0.143	0	0.238	0.19
	40	**Body**	0.238	0.238	0.333	0.238	0.143	0.238	0	0.238
		Headline	0.19	0.238	0.19	0.143	0.19	0.19	0.238	0
Sharpe Ratio	**10**	**Body**	0	0.238	0.286	0.19	0.19	0.19	0.19	0.19
		Headline	0.238	0	0.333	0.238	0.19	0.19	0.286	0.238
	20	**Body**	0.286	0.333	0	0.19	0.238	0.238	0.333	0.19
		Headline	0.19	0.238	0.19	0	0.19	0.143	0.19	0.238
	30	**Body**	0.19	0.19	0.238	0.19	0	0.143	0.143	0.19
		Headline	0.19	0.19	0.238	0.143	0.143	0	0.19	0.19
	40	**Body**	0.19	0.286	0.333	0.19	0.143	0.19	0	0.238
		Headline	0.19	0.238	0.19	0.238	0.19	0.19	0.238	0
MDD	**10**	**Body**	0	0.143	0.19	0.286	0.143	0.286	0.143	0.286
		Headline	0.143	0	0.238	0.19	0.143	0.333	0.143	0.381*
	20	**Body**	0.19	0.238	0	0.143	0.19	0.19	0.238	0.19
		Headline	0.286	0.19	0.143	0	0.238	0.19	0.143	0.286
	30	**Body**	0.143	0.143	0.19	0.238	0	0.238	0.143	0.286
		Headline	0.286	0.333	0.19	0.19	0.238	0	0.238	0.143
	40	**Body**	0.143	0.143	0.238	0.143	0.143	0.238	0	0.286
		Headline	0.286	0.381*	0.19	0.286	0.286	0.143	0.286	0
Calmar Ratio	**10**	**Body**	0	0.143	0.286	0.143	0.143	0.19	0.19	0.143
		Headline	0.143	0	0.333	0.143	0.19	0.238	0.238	0.238
	20	**Body**	0.286	0.333	0	0.19	0.286	0.286	0.333	0.19
		Headline	0.143	0.143	0.19	0	0.19	0.143	0.19	0.238
	30	**Body**	0.143	0.19	0.286	0.19	0	0.19	0.143	0.19
		Headline	0.19	0.238	0.286	0.143	0.19	0	0.19	0.19
	40	**Body**	0.19	0.238	0.333	0.19	0.143	0.19	0	0.238
		Headline	0.143	0.238	0.19	0.238	0.19	0.19	0.238	0

The significance is reported for the following levels: * p<0.1; ** p< 0.05; *** p<0.01.

C Additional Robustness Checks: Idiosyncratic Approaches

In this chapter we show a full comparison of the results for the idiosyncratic approaches depending on the input parameters we use for our models. In particular, in Figure C.1, Figure C.2, Figure C.3, and Figure C.4, we show the distribution of the results respectively when using $n = 10, 20, 30, 40$, where n defines the percentile $P_{n\,\Delta t}$ that we use as a threshold for the sentiment scores for the news we trade on $E(n)_{\Delta t}$ (4.5). In each figure we show how the distribution of the results change when using a different measure of similarity ($C^{\text{median}}_{\Delta t}$, $C^{\text{min}}_{\Delta t}$, and $C^{\text{mean}}_{\Delta t}$) (4.2) and a different m for the percentile $P_{m\,\Delta t}$ that we use as a threshold for the similarity scores for the news we trade on $I(m)_{\Delta t}$ (4.3). As a benchmark, we compare the results of the embedding approach to the results of the syntax and plain RIC approach using the same sentiment percentile $P_{n\,\Delta t}$ (4.5). The results are based on 20 subsamples of news, each consisting of 365 randomly selected non-consecutive days. To check if there is any significant difference in the distribution of the results when using different ms or different similarity measure, we run two-sample Kolmogorov-Smirnof tests. The results of those tests are presented in Table C.1, C.2, C.3, and C.4 respectively.

For $n = 10$ (Figure C.1 and Table C.1), we do not see strongly significant differences in the distribution of the results when using different approaches or different parameters for the embedding approach.

For $n = 20$ (Figure C.2 and Table C.2), we see significant differences in the distribution of the results for most of the evaluation metrics for the embedding approach compared to using the syntax and plain RIC approaches, when using $C^{\text{min}}_{\Delta t}$ as similarity measure between the clauses and the market component and $m = 30$ for the similarity threshold.

For $n = 30$ (Figure C.3 and Table C.3), we find significant differences in the distribution of the results for most of the evaluation metrics between the embedding and the syntax approach when using $m \in \{10, 20\}$ for all the similarity measures ($C^{\text{median}}_{\Delta t}$, $C^{\text{min}}_{\Delta t}$, and $C^{\text{mean}}_{\Delta t}$).

For $n = 40$ (Figure C.4 and Table C.4), there are significant differences in the distribution of the results for most of the evaluation metrics between the embedding and the plain RIC approach when using $m \in \{20, 30\}$ for all the similarity measures ($C^{\text{median}}_{\Delta t}$, $C^{\text{min}}_{\Delta t}$, and $C^{\text{mean}}_{\Delta t}$), and we see significant differences between the syntax and embedding approach for most of the similarity measures and ms.

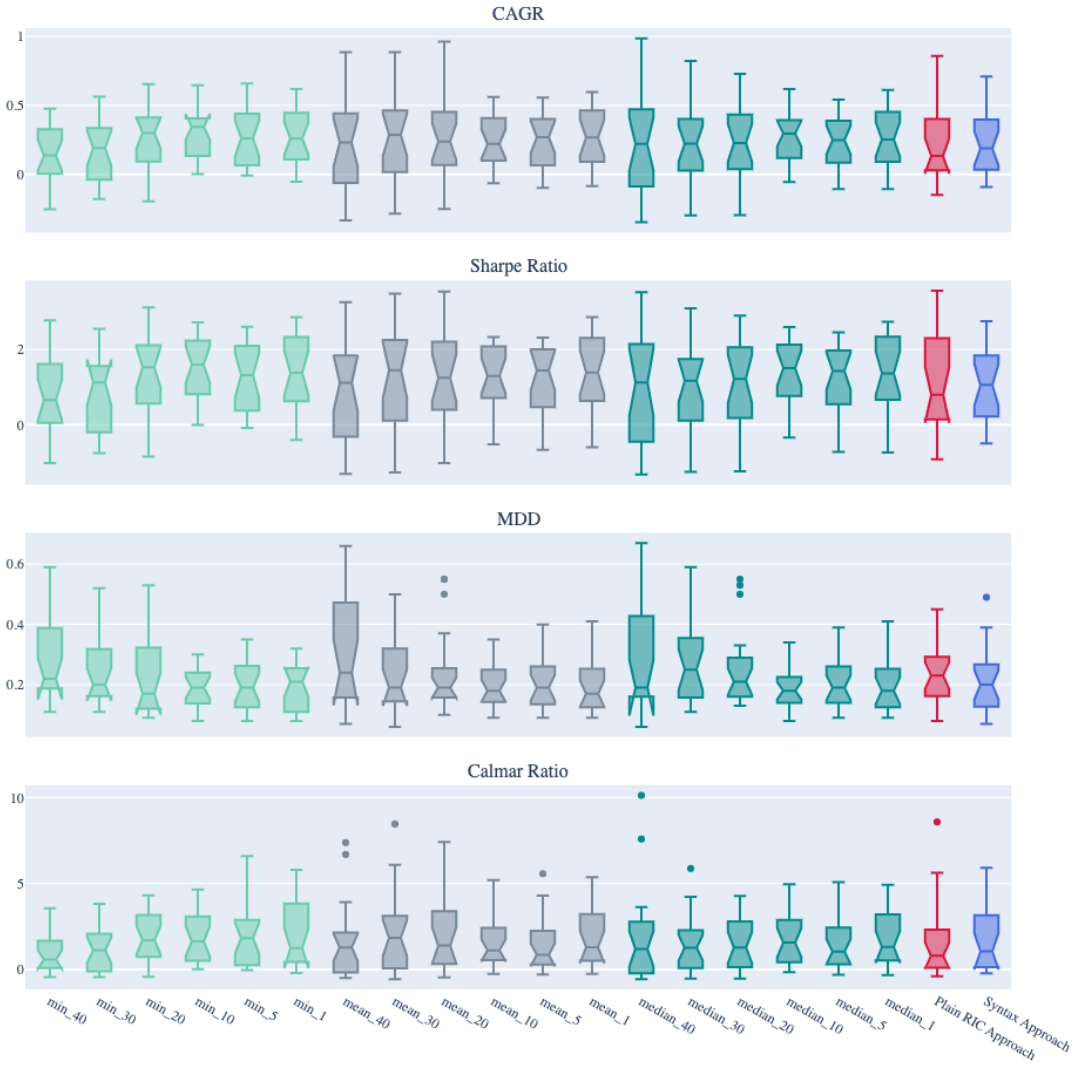

Figure C.1: Sensitivity of the embedding approach to the similarity measures (i.e. $C_{\Delta t}^{\text{median}}$, $C_{\Delta t}^{\text{min}}$, and $C_{\Delta t}^{\text{mean}}$) (4.2), and the m we use for the similarity threshold $P_{m\,\Delta t}$ between idiosyncratic clauses and market component (4.3). We use $\mathbf{n = 10}$ for the percentile $P_{n\,\Delta t}$ that we use as a sentiment threshold for the daily set of news we trade on $E(n)_{\Delta t}$ (4.5). The results shown are based on 20 subsamples of news, each consisting of 365 randomly selected non-consecutive days. In the x-axis we show the combination of similarity measures and ms tested. For example, "mean_1" indicates that we use the measure $C_{\Delta t}^{\text{mean}}$ and $m = 1$. We benchmark the embedding approach to the syntax and the plain RIC approach.

Table C.1: Kolmogorov-Smirnov two-sample test statistics and significance levels for the idiosyncratic approaches using **n=10** for the percentile $P_{n\,\Delta t}$ that we use as sentiment threshold to define the set of news we will trade on (4.5). The results shown are based on 20 sub-samples of news, each consisting of 365 randomly selected non-consecutive days. For the embedding approach, we compare each measure that we use as similarity score between our set of clauses and the daily *market component* set of news (i.e. $C^{\text{median}}_{\Delta t}$, $C^{\text{min}}_{\Delta t}$, and $C^{\text{mean}}_{\Delta t}$) (4.2), and the m we use for the percentile $P_{m\,\Delta t}$ that we use as similarity threshold for our daily set of idiosyncratic news $I(m)_{\Delta t}$ (4.3). We benchmark the embedding approach to the syntax and the plain RIC approach.

			Min						Mean						Median							
Metric	**Measure**	m	**40**	**30**	**20**	**10**	**5**	**1**	**40**	**30**	**20**	**10**	**5**	**1**	**40**	**30**	**20**	**10**	**5**	**1**	**Syntax Approach**	**Plain RIC Approach**
CAGR	**Min**	**40**	0	0.19	0.381*	0.333	0.286	0.333	0.238	0.381*	0.286	0.238	0.238	0.333	0.333	0.238	0.286	0.286	0.286	0.333	0.19	0.238
		30	0.19	0	0.238	0.286	0.286	0.286	0.19	0.333	0.238	0.238	0.238	0.286	0.286	0.19	0.19	0.238	0.238	0.286	0.19	0.238
		20	0.381*	0.238	0	0.238	0.19	0.19	0.19	0.333	0.19	0.19	0.19	0.238	0.238	0.19	0.19	0.143	0.19	0.19	0.286	0.286
		10	0.333	0.286	0.238	0	0.19	0.19	0.333	0.238	0.19	0.19	0.238	0.143	0.381*	0.238	0.238	0.19	0.238	0.143	0.286	0.286
		5	0.286	0.286	0.19	0.19	0	0.19	0.333	0.238	0.19	0.143	0.143	0.143	0.381*	0.19	0.19	0.143	0.095	0.143	0.238	0.143
		1	0.333	0.286	0.19	0.19	0.19	0	0.333	0.238	0.143	0.19	0.238	0.143	0.333	0.19	0.238	0.143	0.19	0.143	0.19	0.19
	Mean	**40**	0.238	0.19	0.19	0.333	0.333	0.333	0	0.286	0.19	0.286	0.286	0.333	0.19	0.19	0.19	0.286	0.286	0.333	0.19	0.286
		30	0.381*	0.333	0.333	0.238	0.238	0.238	0.286	0	0.19	0.238	0.286	0.238	0.333	0.238	0.238	0.286	0.238	0.238	0.286	0.286
		20	0.286	0.238	0.19	0.19	0.19	0.143	0.19	0.19	0	0.143	0.19	0.143	0.238	0.095	0.095	0.143	0.19	0.143	0.286	0.19
		10	0.238	0.238	0.19	0.19	0.143	0.19	0.286	0.238	0.143	0	0.19	0.143	0.333	0.143	0.143	0.143	0.095	0.19	0.238	0.19
		5	0.238	0.238	0.19	0.238	0.143	0.238	0.286	0.286	0.19	0.19	0	0.238	0.333	0.143	0.143	0.19	0.143	0.238	0.19	0.19
		1	0.333	0.286	0.238	0.143	0.143	0.143	0.333	0.238	0.143	0.143	0.238	0	0.333	0.19	0.238	0.19	0.19	0.095	0.238	0.19
	Median	**40**	0.333	0.286	0.238	0.381*	0.381*	0.333	0.19	0.333	0.238	0.333	0.333	0.333	0	0.238	0.238	0.333	0.333	0.333	0.19	0.333
		30	0.238	0.19	0.19	0.238	0.19	0.19	0.19	0.238	0.095	0.143	0.143	0.19	0.238	0	0.095	0.143	0.19	0.19	0.19	0.143
		20	0.286	0.19	0.19	0.238	0.19	0.238	0.19	0.238	0.095	0.143	0.143	0.238	0.238	0.095	0	0.19	0.19	0.19	0.238	0.143
		10	0.286	0.238	0.143	0.19	0.143	0.143	0.286	0.286	0.143	0.143	0.19	0.19	0.333	0.143	0.19	0	0.143	0.143	0.238	0.238
		5	0.286	0.238	0.19	0.238	0.095	0.19	0.286	0.238	0.19	0.095	0.143	0.19	0.333	0.19	0.19	0.143	0	0.19	0.19	0.19
		1	0.333	0.286	0.19	0.143	0.143	0.143	0.333	0.238	0.143	0.19	0.238	0.095	0.333	0.19	0.19	0.143	0.19	0	0.238	0.19
	Syntax Approach		0.19	0.19	0.286	0.286	0.238	0.19	0.19	0.286	0.286	0.238	0.19	0.238	0.19	0.19	0.238	0.238	0.19	0.238	0	0.19
	Plain RIC Approach		0.238	0.238	0.286	0.286	0.143	0.19	0.286	0.286	0.19	0.19	0.19	0.19	0.333	0.143	0.143	0.238	0.19	0.19	0.19	0
Sharpe Ratio	**Min**	**40**	0	0.238	0.333	0.333	0.286	0.333	0.19	0.333	0.286	0.333	0.238	0.333	0.333	0.238	0.238	0.333	0.238	0.333	0.238	0.238
		30	0.238	0	0.381*	0.333	0.286	0.286	0.143	0.381*	0.238	0.286	0.238	0.286	0.333	0.19	0.19	0.333	0.238	0.286	0.19	0.19
		20	0.333	0.381*	0	0.238	0.19	0.238	0.286	0.19	0.238	0.143	0.143	0.286	0.238	0.286	0.238	0.143	0.143	0.19	0.333	0.238
		10	0.333	0.333	0.238	0	0.143	0.19	0.333	0.238	0.19	0.238	0.238	0.143	0.381*	0.286	0.238	0.143	0.19	0.143	0.333	0.19
		5	0.286	0.286	0.19	0.143	0	0.19	0.333	0.19	0.19	0.19	0.19	0.19	0.381*	0.19	0.19	0.143	0.143	0.143	0.238	0.143
		1	0.333	0.286	0.238	0.19	0.19	0	0.333	0.19	0.143	0.286	0.286	0.095	0.333	0.19	0.238	0.143	0.286	0.095	0.238	0.19
	Mean	**40**	0.19	0.143	0.286	0.333	0.333	0.333	0	0.286	0.19	0.238	0.286	0.333	0.19	0.19	0.19	0.238	0.286	0.333	0.143	0.238
		30	0.333	0.381*	0.19	0.238	0.19	0.19	0.286	0	0.238	0.238	0.238	0.238	0.333	0.238	0.19	0.143	0.19	0.238	0.286	0.238
		20	0.286	0.238	0.238	0.19	0.19	0.143	0.19	0.238	0	0.238	0.19	0.143	0.238	0.143	0.143	0.19	0.19	0.143	0.238	0.143
		10	0.333	0.286	0.143	0.238	0.19	0.286	0.238	0.238	0.238	0	0.19	0.333	0.286	0.143	0.143	0.19	0.143	0.238	0.286	0.143
		5	0.238	0.238	0.143	0.238	0.19	0.286	0.286	0.238	0.19	0.19	0	0.238	0.286	0.19	0.143	0.19	0.143	0.238	0.238	0.143
		1	0.333	0.286	0.286	0.143	0.19	0.095	0.333	0.238	0.143	0.333	0.238	0	0.333	0.238	0.238	0.143	0.238	0.095	0.286	0.19
	Median	**40**	0.333	0.333	0.238	0.381*	0.381*	0.333	0.19	0.333	0.238	0.286	0.286	0.333	0	0.238	0.286	0.333	0.333	0.333	0.19	0.333
		30	0.238	0.19	0.286	0.286	0.19	0.19	0.19	0.238	0.143	0.143	0.19	0.238	0.238	0	0.095	0.19	0.19	0.238	0.19	0.095
		20	0.238	0.19	0.238	0.238	0.19	0.238	0.19	0.19	0.143	0.143	0.143	0.238	0.286	0.095	0	0.19	0.143	0.19	0.19	0.095
		10	0.333	0.333	0.143	0.143	0.143	0.143	0.238	0.143	0.19	0.19	0.19	0.143	0.333	0.19	0.19	0	0.19	0.19	0.286	0.19
		5	0.238	0.238	0.143	0.19	0.143	0.286	0.286	0.19	0.19	0.143	0.143	0.238	0.333	0.19	0.143	0.19	0	0.238	0.238	0.143
		1	0.333	0.286	0.19	0.143	0.143	0.095	0.333	0.238	0.143	0.238	0.238	0.095	0.333	0.238	0.19	0.19	0.238	0	0.238	0.19
	Syntax Approach		0.238	0.19	0.333	0.333	0.238	0.238	0.143	0.286	0.238	0.286	0.238	0.286	0.19	0.19	0.19	0.286	0.238	0.238	0	0.19
	Plain RIC Approach		0.238	0.19	0.238	0.19	0.143	0.19	0.238	0.238	0.143	0.143	0.143	0.19	0.333	0.095	0.095	0.19	0.143	0.19	0.19	0
MDD	**Min**	**40**	0	0.143	0.333	0.333	0.286	0.333	0.143	0.238	0.238	0.286	0.238	0.333	0.19	0.143	0.19	0.333	0.286	0.286	0.19	0.286
		30	0.143	0	0.238	0.238	0.238	0.238	0.19	0.143	0.19	0.238	0.19	0.238	0.19	0.19	0.143	0.238	0.238	0.19	0.19	0.238
		20	0.333	0.238	0	0.286	0.238	0.238	0.286	0.19	0.238	0.19	0.143	0.143	0.286	0.286	0.333	0.19	0.19	0.143	0.238	0.19
		10	0.333	0.238	0.286	0	0.238	0.19	0.381*	0.286	0.238	0.095	0.19	0.19	0.333	0.333	0.238	0.143	0.19	0.19	0.238	0.143
		5	0.286	0.238	0.238	0.238	0	0.095	0.333	0.238	0.238	0.19	0.143	0.143	0.286	0.286	0.238	0.238	0.143	0.143	0.19	0.143
		1	0.333	0.238	0.238	0.19	0.095	0	0.381*	0.19	0.238	0.143	0.19	0.19	0.333	0.286	0.286	0.19	0.143	0.143	0.19	0.19
	Mean	**40**	0.143	0.19	0.286	0.381*	0.333	0.381*	0	0.238	0.286	0.381*	0.286	0.286	0.143	0.143	0.238	0.333	0.286	0.286	0.286	0.286
		30	0.238	0.143	0.19	0.286	0.238	0.19	0.238	0	0.095	0.19	0.143	0.143	0.238	0.238	0.19	0.19	0.143	0.143	0.19	0.19
		20	0.238	0.19	0.238	0.238	0.238	0.238	0.286	0.095	0	0.19	0.143	0.238	0.19	0.286	0.238	0.19	0.19	0.19	0.238	0.19
		10	0.286	0.238	0.19	0.095	0.19	0.143	0.381*	0.19	0.19	0	0.143	0.19	0.286	0.286	0.238	0.143	0.143	0.143	0.19	0.143
		5	0.238	0.19	0.143	0.19	0.143	0.19	0.286	0.143	0.143	0.143	0	0.143	0.238	0.238	0.238	0.143	0.048	0.095	0.143	0.143
		1	0.333	0.238	0.143	0.19	0.143	0.19	0.286	0.143	0.238	0.19	0.143	0	0.286	0.238	0.286	0.19	0.143	0.048	0.238	0.143
	Median	**40**	0.19	0.19	0.286	0.333	0.286	0.333	0.143	0.238	0.19	0.286	0.238	0.286	0	0.143	0.19	0.286	0.286	0.238	0.238	0.286
		30	0.143	0.19	0.286	0.333	0.286	0.286	0.143	0.238	0.286	0.286	0.238	0.238	0.143	0	0.143	0.333	0.238	0.238	0.143	0.238
		20	0.19	0.143	0.333	0.238	0.238	0.286	0.238	0.19	0.238	0.238	0.238	0.286	0.19	0.143	0	0.238	0.19	0.238	0.143	0.286
		10	0.333	0.238	0.19	0.143	0.238	0.19	0.333	0.19	0.19	0.143	0.143	0.19	0.286	0.333	0.238	0	0.19	0.143	0.286	0.143
		5	0.286	0.238	0.19	0.19	0.143	0.143	0.286	0.143	0.19	0.143	0.048	0.143	0.286	0.238	0.19	0.19	0	0.143	0.19	0.19
		1	0.286	0.19	0.143	0.19	0.143	0.143	0.286	0.143	0.19	0.143	0.095	0.048	0.238	0.238	0.238	0.143	0.143	0	0.19	0.143
	Syntax Approach		0.19	0.19	0.238	0.238	0.19	0.19	0.286	0.19	0.238	0.19	0.143	0.238	0.238	0.143	0.143	0.286	0.19	0.19	0	0.238
	Plain RIC Approach		0.286	0.238	0.19	0.143	0.143	0.19	0.286	0.19	0.19	0.143	0.143	0.143	0.286	0.238	0.286	0.143	0.19	0.143	0.238	0
Calmar Ratio	**Min**	**40**	0	0.19	0.381*	0.333	0.286	0.286	0.19	0.333	0.286	0.286	0.286	0.333	0.286	0.19	0.238	0.333	0.286	0.333	0.19	0.238
		30	0.19	0	0.286	0.286	0.286	0.286	0.19	0.286	0.238	0.19	0.238	0.286	0.238	0.143	0.19	0.238	0.238	0.286	0.143	0.19
		20	0.381*	0.286	0	0.238	0.238	0.19	0.286	0.143	0.19	0.143	0.286	0.238	0.238	0.19	0.238	0.143	0.238	0.238	0.333	0.238
		10	0.333	0.286	0.238	0	0.19	0.143	0.333	0.238	0.19	0.19	0.238	0.143	0.381*	0.238	0.238	0.143	0.238	0.143	0.333	0.238
		5	0.286	0.286	0.238	0.19	0	0.19	0.333	0.238	0.19	0.143	0.143	0.143	0.381*	0.19	0.19	0.143	0.19	0.143	0.238	0.19
		1	0.286	0.286	0.19	0.143	0.19	0	0.333	0.238	0.19	0.19	0.143	0.143	0.333	0.19	0.238	0.238	0.238	0.143	0.19	0.19
	Mean	**40**	0.19	0.19	0.286	0.333	0.333	0.333	0	0.286	0.19	0.238	0.286	0.333	0.19	0.19	0.19	0.238	0.286	0.333	0.238	0.238
		30	0.333	0.286	0.143	0.238	0.238	0.238	0.286	0	0.19	0.19	0.238	0.19	0.333	0.19	0.238	0.143	0.19	0.19	0.286	0.238
		20	0.286	0.238	0.19	0.19	0.19	0.19	0.19	0.19	0	0.19	0.19	0.19	0.238	0.19	0.19	0.143	0.143	0.19	0.286	0.19
		10	0.286	0.19	0.143	0.19	0.143	0.19	0.238	0.19	0.19	0	0.19	0.143	0.286	0.143	0.143	0.19	0.143	0.143	0.19	0.143
		5	0.286	0.238	0.286	0.238	0.143	0.143	0.286	0.238	0.19	0.19	0	0.19	0.286	0.143	0.143	0.238	0.19	0.238	0.19	0.143
		1	0.333	0.286	0.238	0.143	0.143	0.143	0.333	0.19	0.19	0.143	0.19	0	0.333	0.19	0.238	0.19	0.19	0.048	0.238	0.19
	Median	**40**	0.286	0.238	0.238	0.381*	0.381*	0.333	0.19	0.333	0.238	0.286	0.286	0.333	0	0.238	0.286	0.333	0.333	0.333	0.19	0.333
		30	0.19	0.143	0.19	0.238	0.19	0.19	0.19	0.19	0.19	0.143	0.143	0.19	0.238	0	0.095	0.19	0.19	0.19	0.19	0.143
		20	0.238	0.19	0.238	0.238	0.19	0.238	0.19	0.238	0.19	0.143	0.143	0.238	0.286	0.095	0	0.19	0.143	0.238	0.238	0.19
		10	0.333	0.238	0.143	0.143	0.143	0.238	0.238	0.143	0.143	0.19	0.238	0.19	0.333	0.19	0.19	0	0.19	0.19	0.333	0.238
		5	0.286	0.238	0.238	0.238	0.19	0.238	0.286	0.19	0.143	0.143	0.19	0.19	0.333	0.19	0.143	0.19	0	0.19	0.19	0.19
		1	0.333	0.286	0.238	0.143	0.143	0.143	0.333	0.19	0.19	0.143	0.238	0.048	0.333	0.19	0.238	0.19	0.19	0	0.238	0.19
	Syntax Approach		0.19	0.143	0.333	0.333	0.238	0.19	0.238	0.286	0.286	0.19	0.19	0.238	0.19	0.19	0.238	0.333	0.19	0.238	0	0.19
	Plain RIC Approach		0.238	0.19	0.238	0.238	0.19	0.19	0.238	0.238	0.19	0.143	0.143	0.19	0.333	0.143	0.19	0.238	0.19	0.19	0.19	0

The significance is reported for the following levels: * p<0.1; ** p< 0.05; *** p<0.01.

Figure C.2: Sensitivity of the embedding approach to the similarity measures (i.e. $C_{\Delta t}^{\text{median}}$, $C_{\Delta t}^{\text{min}}$, and $C_{\Delta t}^{\text{mean}}$) (4.2), and the m we use for the similarity threshold $P_{m\,\Delta t}$ between idiosyncratic clauses and market component (4.3). We use $\mathbf{n = 20}$ for the percentile $P_{n\,\Delta t}$ that we use as a sentiment threshold for the daily set of news we trade on $E(n)_{\Delta t}$ (4.5). The results shown are based on 20 subsamples of news, each consisting of 365 randomly selected non-consecutive days. In the x-axis we show the combination of similarity measures and ms tested. For example, "mean_1" indicates that we use the measure $C_{\Delta t}^{\text{mean}}$ and $m = 1$. We benchmark the embedding approach to the syntax and the plain RIC approach.

Table C.2: Kolmogorov-Smirnov two-sample test statistics and significance levels for the idiosyncratic approaches using **n=20** for the percentile $P_{n\,\Delta t}$ that we use as sentiment threshold to define the set of news we will trade on (4.5). The results shown are based on 20 sub-samples of news, each consisting of 365 randomly selected non-consecutive days. For the embedding approach, we compare each measure that we use as similarity score between our set of clauses and the daily *market component* set of news (i.e. $C^{\text{median}}_{\Delta t}$, $C^{\text{min}}_{\Delta t}$, and $C^{\text{mean}}_{\Delta t}$) (4.2), and the m we use for the percentile $P_{m\,\Delta t}$ that we use as similarity threshold for our daily set of idiosyncratic news $I(m)_{\Delta t}$ (4.3). We benchmark the embedding approach to the syntax and the plain RIC approach.

			Min						Mean						Median							
Metric	**Measure**	m	**40**	**30**	**20**	**10**	**5**	**1**	**40**	**30**	**20**	**10**	**5**	**1**	**40**	**30**	**20**	**10**	**5**	**1**	**Syntax Approach**	**Plain RIC Approach**
		40	0	0.238	0.333	0.476**	0.429**	0.429**	0.286	0.381*	0.381*	0.333	0.381*	0.381*	0.381*	0.381*	0.381*	0.429**	0.333	0.429**	0.286	0.381*
		30	0.238	0	0.381*	0.524***	0.333	0.333	0.333	0.429**	0.429**	0.333	0.381*	0.333	0.429**	0.381*	0.429**	0.476**	0.381*	0.333	0.476**	0.429**
		20	0.333	0.381*	0	0.238	0.238	0.238	0.238	0.286	0.286	0.143	0.143	0.143	0.333	0.238	0.286	0.238	0.143	0.19	0.238	0.286
	Min	**10**	0.476**	0.524***	0.238	0	0.429**	0.381*	0.381*	0.286	0.238	0.333	0.238	0.286	0.286	0.238	0.238	0.238	0.238	0.238	0.19	0.238
		5	0.429**	0.333	0.238	0.429**	0	0.19	0.286	0.333	0.333	0.19	0.238	0.286	0.333	0.238	0.286	0.333	0.238	0.238	0.333	0.286
		1	0.429**	0.333	0.238	0.381*	0.19	0	0.333	0.286	0.333	0.19	0.238	0.19	0.333	0.238	0.333	0.238	0.238	0.19	0.238	0.286
		40	0.286	0.333	0.238	0.381*	0.286	0.333	0	0.143	0.286	0.238	0.286	0.286	0.143	0.143	0.238	0.286	0.238	0.333	0.19	0.238
		30	0.381*	0.429**	0.286	0.286	0.333	0.286	0.143	0	0.19	0.238	0.286	0.333	0.095	0.095	0.143	0.238	0.286	0.333	0.143	0.143
		20	0.381*	0.429**	0.286	0.238	0.333	0.333	0.286	0.19	0	0.19	0.286	0.286	0.19	0.19	0.143	0.19	0.238	0.286	0.143	0.143
	Mean	**10**	0.333	0.333	0.143	0.333	0.19	0.19	0.238	0.238	0.19	0	0.143	0.143	0.19	0.19	0.19	0.238	0.143	0.143	0.19	0.19
		5	0.381*	0.381*	0.143	0.238	0.238	0.238	0.286	0.286	0.286	0.143	0	0.095	0.286	0.238	0.286	0.238	0.095	0.143	0.19	0.238
		1	0.381*	0.333	0.143	0.286	0.286	0.19	0.286	0.333	0.286	0.143	0.095	0	0.286	0.286	0.286	0.238	0.143	0.095	0.19	0.286
CAGR		**40**	0.381*	0.429**	0.333	0.286	0.333	0.333	0.143	0.095	0.19	0.19	0.286	0.286	0	0.143	0.143	0.19	0.286	0.286	0.143	0.19
		30	0.381*	0.381*	0.238	0.238	0.238	0.238	0.143	0.095	0.19	0.19	0.238	0.286	0.143	0	0.19	0.19	0.238	0.286	0.143	0.19
		20	0.381*	0.429**	0.286	0.238	0.286	0.333	0.238	0.143	0.143	0.19	0.286	0.286	0.143	0.19	0	0.19	0.286	0.286	0.143	0.19
	Median	**10**	0.429**	0.476**	0.238	0.238	0.333	0.238	0.286	0.238	0.19	0.238	0.238	0.238	0.19	0.19	0.19	0	0.143	0.19	0.143	0.19
		5	0.333	0.381*	0.143	0.238	0.238	0.238	0.238	0.286	0.238	0.143	0.095	0.143	0.286	0.238	0.286	0.143	0	0.143	0.19	0.238
		1	0.429**	0.333	0.19	0.238	0.238	0.19	0.333	0.333	0.286	0.143	0.143	0.095	0.286	0.286	0.286	0.19	0.143	0	0.238	0.286
	Syntax Approach		0.286	0.476**	0.238	0.19	0.333	0.238	0.19	0.143	0.143	0.19	0.19	0.19	0.143	0.143	0.143	0.143	0.19	0.238	0	0.143
	Plain RIC Approach		0.381*	0.429**	0.286	0.238	0.286	0.286	0.238	0.143	0.143	0.19	0.238	0.286	0.19	0.19	0.19	0.19	0.238	0.286	0.143	0
		40	0	0.238	0.333	0.476**	0.429**	0.429**	0.286	0.381*	0.381*	0.333	0.381*	0.429**	0.333	0.286	0.333	0.429**	0.333	0.429**	0.286	0.381*
		30	0.238	0	0.429**	0.619***	0.333	0.333	0.286	0.429**	0.429**	0.333	0.476**	0.381*	0.429**	0.333	0.476**	0.571***	0.476**	0.429**	0.524***	0.476**
		20	0.333	0.429**	0	0.286	0.238	0.238	0.238	0.238	0.286	0.095	0.095	0.143	0.238	0.238	0.238	0.19	0.143	0.143	0.238	0.286
	Min	**10**	0.476**	0.619***	0.286	0	0.429**	0.333	0.429**	0.381*	0.286	0.333	0.238	0.286	0.286	0.333	0.238	0.19	0.238	0.286	0.19	0.238
		5	0.429**	0.333	0.238	0.429**	0	0.19	0.286	0.333	0.333	0.238	0.238	0.19	0.333	0.238	0.381*	0.333	0.286	0.238	0.286	0.333
		1	0.429**	0.333	0.238	0.333	0.19	0	0.286	0.333	0.333	0.238	0.19	0.19	0.333	0.238	0.381*	0.286	0.238	0.19	0.286	0.333
		40	0.286	0.286	0.238	0.429**	0.286	0.286	0	0.19	0.238	0.238	0.286	0.286	0.19	0.143	0.238	0.286	0.238	0.286	0.238	0.286
		30	0.381*	0.429**	0.238	0.381*	0.333	0.333	0.19	0	0.19	0.19	0.238	0.238	0.095	0.143	0.143	0.238	0.19	0.238	0.19	0.238
		20	0.381*	0.429**	0.286	0.286	0.333	0.333	0.238	0.19	0	0.238	0.238	0.238	0.19	0.19	0.19	0.19	0.238	0.286	0.143	0.19
	Mean	**10**	0.333	0.333	0.095	0.333	0.238	0.238	0.238	0.19	0.238	0	0.143	0.143	0.19	0.19	0.238	0.238	0.143	0.095	0.238	0.19
Sharpe Ratio		**5**	0.381*	0.476**	0.095	0.238	0.238	0.19	0.286	0.238	0.238	0.143	0	0.095	0.238	0.19	0.286	0.238	0.143	0.095	0.19	0.238
		1	0.429**	0.381*	0.143	0.286	0.19	0.19	0.286	0.238	0.238	0.143	0.095	0	0.238	0.238	0.238	0.238	0.143	0.095	0.19	0.286
		40	0.333	0.429**	0.238	0.286	0.333	0.333	0.19	0.095	0.19	0.19	0.238	0.238	0	0.143	0.143	0.19	0.19	0.238	0.095	0.143
		30	0.286	0.333	0.238	0.333	0.238	0.238	0.143	0.143	0.19	0.19	0.19	0.238	0.143	0	0.19	0.238	0.19	0.19	0.19	0.19
		20	0.333	0.476**	0.238	0.238	0.381*	0.381*	0.238	0.143	0.19	0.238	0.286	0.238	0.143	0.19	0	0.19	0.238	0.238	0.19	0.19
	Median	**10**	0.429**	0.571***	0.19	0.19	0.333	0.286	0.286	0.238	0.19	0.238	0.238	0.238	0.19	0.238	0.19	0	0.143	0.19	0.19	0.19
		5	0.333	0.476**	0.143	0.238	0.286	0.238	0.238	0.19	0.238	0.143	0.143	0.143	0.19	0.19	0.238	0.143	0	0.143	0.19	0.238
		1	0.429**	0.429**	0.143	0.286	0.238	0.19	0.286	0.238	0.286	0.095	0.095	0.095	0.238	0.19	0.238	0.19	0.143	0	0.238	0.286
	Syntax Approach		0.286	0.524***	0.238	0.19	0.286	0.286	0.238	0.19	0.143	0.238	0.19	0.19	0.095	0.19	0.19	0.19	0.19	0.238	0	0.143
	Plain RIC Approach		0.381*	0.476**	0.286	0.238	0.333	0.333	0.286	0.238	0.19	0.19	0.238	0.286	0.143	0.19	0.19	0.19	0.238	0.286	0.143	0
		40	0	0.095	0.238	0.286	0.286	0.238	0.143	0.095	0.333	0.19	0.19	0.238	0.095	0.143	0.333	0.238	0.238	0.238	0.333	0.286
		30	0.095	0	0.238	0.333	0.238	0.286	0.19	0.143	0.381*	0.238	0.19	0.238	0.095	0.143	0.333	0.286	0.286	0.286	0.381*	0.333
		20	0.238	0.238	0	0.19	0.238	0.286	0.333	0.19	0.238	0.095	0.095	0.19	0.286	0.286	0.095	0.143	0.143	0.143	0.238	0.19
	Min	**10**	0.286	0.333	0.19	0	0.238	0.143	0.381*	0.238	0.286	0.19	0.143	0.19	0.333	0.238	0.238	0.19	0.19	0.19	0.143	0.238
		5	0.286	0.238	0.238	0.238	0	0.19	0.333	0.19	0.333	0.19	0.19	0.238	0.333	0.19	0.333	0.238	0.19	0.238	0.333	0.286
		1	0.238	0.286	0.286	0.143	0.19	0	0.333	0.19	0.333	0.238	0.19	0.286	0.286	0.19	0.333	0.238	0.238	0.238	0.19	0.286
		40	0.143	0.19	0.333	0.381*	0.333	0.333	0	0.19	0.429**	0.286	0.286	0.333	0.143	0.19	0.381*	0.286	0.286	0.333	0.476**	0.381*
		30	0.095	0.143	0.19	0.238	0.19	0.19	0.19	0	0.333	0.238	0.19	0.19	0.19	0.095	0.286	0.238	0.238	0.238	0.286	0.286
		20	0.333	0.381*	0.238	0.286	0.333	0.333	0.429**	0.333	0	0.19	0.19	0.19	0.381*	0.381*	0.19	0.143	0.143	0.143	0.286	0.19
	Mean	**10**	0.19	0.238	0.095	0.19	0.19	0.238	0.286	0.238	0.19	0	0.095	0.143	0.238	0.238	0.19	0.095	0.095	0.095	0.19	0.143
MDD		**5**	0.19	0.19	0.095	0.143	0.19	0.19	0.286	0.19	0.19	0.095	0	0.143	0.238	0.238	0.143	0.095	0.095	0.095	0.19	0.143
		1	0.238	0.238	0.19	0.19	0.238	0.286	0.333	0.19	0.19	0.143	0.143	0	0.286	0.286	0.143	0.143	0.143	0.095	0.238	0.19
		40	0.095	0.095	0.286	0.333	0.333	0.286	0.143	0.19	0.381*	0.238	0.238	0.286	0	0.19	0.333	0.286	0.286	0.286	0.381*	0.333
		30	0.143	0.143	0.286	0.238	0.19	0.19	0.19	0.095	0.381*	0.238	0.238	0.286	0.19	0	0.381*	0.286	0.238	0.286	0.333	0.333
		20	0.333	0.333	0.095	0.238	0.333	0.333	0.381*	0.286	0.19	0.19	0.143	0.143	0.333	0.381*	0	0.143	0.143	0.095	0.286	0.143
	Median	**10**	0.238	0.286	0.143	0.19	0.238	0.238	0.286	0.238	0.143	0.095	0.095	0.143	0.286	0.286	0.143	0	0.095	0.143	0.19	0.143
		5	0.238	0.286	0.143	0.19	0.19	0.238	0.286	0.238	0.143	0.095	0.095	0.143	0.286	0.238	0.143	0.095	0	0.095	0.19	0.143
		1	0.238	0.286	0.143	0.19	0.238	0.238	0.333	0.238	0.143	0.095	0.095	0.095	0.286	0.286	0.095	0.143	0.095	0	0.19	0.19
	Syntax Approach		0.333	0.381*	0.238	0.143	0.333	0.19	0.476**	0.286	0.286	0.19	0.19	0.238	0.381*	0.333	0.286	0.19	0.19	0.19	0	0.19
	Plain RIC Approach		0.286	0.333	0.19	0.238	0.286	0.286	0.381*	0.286	0.19	0.143	0.143	0.19	0.333	0.333	0.143	0.143	0.143	0.19	0.19	0
		40	0	0.238	0.333	0.476**	0.429**	0.429**	0.238	0.286	0.381*	0.333	0.333	0.429**	0.286	0.286	0.381*	0.429**	0.381*	0.381*	0.286	0.333
		30	0.238	0	0.381*	0.619***	0.333	0.429**	0.286	0.333	0.381*	0.381*	0.429**	0.381*	0.381*	0.381*	0.476**	0.524***	0.429**	0.381*	0.524***	0.524***
		20	0.333	0.381*	0	0.333	0.238	0.19	0.238	0.143	0.286	0.095	0.19	0.143	0.19	0.143	0.238	0.19	0.19	0.19	0.19	0.238
	Min	**10**	0.476**	0.619***	0.333	0	0.476**	0.429**	0.381*	0.333	0.333	0.286	0.286	0.286	0.286	0.333	0.238	0.238	0.238	0.286	0.238	0.238
		5	0.429**	0.333	0.238	0.476**	0	0.19	0.286	0.238	0.381*	0.238	0.333	0.238	0.286	0.19	0.333	0.381*	0.286	0.286	0.333	0.381*
		1	0.429**	0.429**	0.19	0.429**	0.19	0	0.333	0.286	0.333	0.238	0.286	0.238	0.286	0.238	0.429**	0.333	0.286	0.238	0.286	0.286
		40	0.238	0.286	0.238	0.381*	0.286	0.333	0	0.19	0.286	0.238	0.286	0.286	0.143	0.19	0.333	0.286	0.238	0.333	0.286	0.286
		30	0.286	0.333	0.143	0.333	0.238	0.286	0.19	0	0.238	0.19	0.19	0.238	0.095	0.143	0.143	0.238	0.238	0.238	0.19	0.238
		20	0.381*	0.381*	0.286	0.333	0.381*	0.333	0.286	0.238	0	0.238	0.333	0.286	0.19	0.238	0.143	0.286	0.286	0.333	0.19	0.19
	Mean	**10**	0.333	0.381*	0.095	0.286	0.238	0.238	0.238	0.19	0.238	0	0.143	0.143	0.19	0.143	0.19	0.143	0.143	0.143	0.143	0.19
Calmar Ratio		**5**	0.333	0.429**	0.19	0.286	0.333	0.286	0.286	0.19	0.333	0.143	0	0.095	0.238	0.19	0.238	0.19	0.19	0.095	0.19	0.19
		1	0.429**	0.381*	0.143	0.286	0.238	0.238	0.286	0.238	0.286	0.143	0.095	0	0.238	0.238	0.238	0.143	0.143	0.143	0.19	0.19
		40	0.286	0.381*	0.19	0.286	0.286	0.286	0.143	0.095	0.19	0.19	0.238	0.238	0	0.095	0.19	0.238	0.238	0.238	0.143	0.143
		30	0.286	0.381*	0.143	0.333	0.19	0.238	0.19	0.143	0.238	0.143	0.19	0.238	0.095	0	0.238	0.19	0.19	0.19	0.19	0.19
		20	0.381*	0.476**	0.238	0.238	0.333	0.429**	0.333	0.143	0.143	0.19	0.238	0.238	0.19	0.238	0	0.19	0.19	0.19	0.143	0.19
		10	0.429**	0.524***	0.19	0.238	0.381*	0.333	0.286	0.238	0.286	0.143	0.19	0.143	0.238	0.19	0.19	0	0.19	0.19	0.143	0.19
		5	0.381*	0.429**	0.19	0.238	0.286	0.286	0.238	0.238	0.286	0.143	0.19	0.143	0.238	0.19	0.19	0.19	0	0.143	0.238	0.238
		1	0.381*	0.381*	0.19	0.286	0.286	0.238	0.333	0.238	0.333	0.143	0.095	0.143	0.238	0.19	0.19	0.19	0.143	0	0.238	0.238
	Syntax Approach		0.286	0.524***	0.19	0.238	0.333	0.286	0.286	0.19	0.19	0.143	0.19	0.19	0.143	0.19	0.143	0.143	0.238	0.238	0	0.143
	Plain RIC Approach		0.333	0.524***	0.238	0.238	0.381*	0.286	0.286	0.238	0.19	0.19	0.19	0.19	0.143	0.19	0.19	0.19	0.238	0.238	0.143	0

The significance is reported for the following levels: * p<0.1; ** p< 0.05; *** p<0.01.

Figure C.3: Sensitivity of the embedding approach to the similarity measures (i.e. $C_{\Delta t}^{\text{median}}$, $C_{\Delta t}^{\text{min}}$, and $C_{\Delta t}^{\text{mean}}$) (4.2), and the m we use for the similarity threshold $P_{m\,\Delta t}$ between idiosyncratic clauses and market component (4.3). We use $\mathbf{n = 30}$ for the percentile $P_{n\,\Delta t}$ that we use as a sentiment threshold for the daily set of news we trade on $E(n)_{\Delta t}$ (4.5). The results shown are based on 20 subsamples of news, each consisting of 365 randomly selected non-consecutive days. In the x-axis we show the combination of similarity measures and ms tested. For example, "mean_1" indicates that we use the measure $C_{\Delta t}^{\text{mean}}$ and $m = 1$. We benchmark the embedding approach to the syntax and the plain RIC approach.

Table C.3: Kolmogorov-Smirnov two-sample test statistics and significance levels for the idiosyncratic approaches using **n=30** for the percentile $P_{n\,\Delta t}$ that we use as sentiment threshold to define the set of news we will trade on (4.5). The results shown are based on 20 sub-samples of news, each consisting of 365 randomly selected non-consecutive days. For the embedding approach, we compare each measure that we use as similarity score between our set of clauses and the daily *market component* set of news (i.e. $C^{\text{median}}_{\Delta t}$, $C^{\text{min}}_{\Delta t}$, and $C^{\text{mean}}_{\Delta t}$) (4.2), and the m we use for the percentile $P_{m\,\Delta t}$ that we use as similarity threshold for our daily set of idiosyncratic news $I(m)_{\Delta t}$ (4.3). We benchmark the embedding approach to the syntax and the plain RIC approach.

Metric	Measure	m	Min						Mean						Median						Syntax Approach	Plain RIC Approach
			40	30	20	10	5	1	40	30	20	10	5	1	40	30	20	10	5	1		
CAGR	Min	40	0	0.19	0.429**	0.429**	0.286	0.286	0.333	0.524***	0.476**	0.429**	0.476**	0.333	0.381*	0.619***	0.429**	0.476**	0.524***	0.333	0.19	0.238
		30	0.19	0	0.333	0.333	0.143	0.19	0.286	0.429**	0.429**	0.333	0.381*	0.286	0.333	0.524***	0.381*	0.429**	0.429**	0.238	0.286	0.19
		20	0.429**	0.333	0	0.19	0.286	0.286	0.19	0.238	0.333	0.19	0.238	0.19	0.238	0.429**	0.286	0.238	0.238	0.19	0.524***	0.429**
		10	0.429**	0.333	0.19	0	0.286	0.286	0.19	0.143	0.238	0.143	0.143	0.19	0.286	0.333	0.238	0.238	0.19	0.19	0.476**	0.333
		5	0.286	0.143	0.286	0.286	0	0.095	0.19	0.286	0.286	0.238	0.333	0.143	0.238	0.381*	0.238	0.286	0.333	0.143	0.333	0.19
		1	0.286	0.19	0.286	0.286	0.095	0	0.19	0.286	0.286	0.238	0.238	0.19	0.238	0.381*	0.238	0.238	0.286	0.19	0.286	0.143
	Mean	40	0.333	0.286	0.19	0.19	0.19	0.19	0	0.238	0.238	0.19	0.19	0.143	0.143	0.381*	0.19	0.19	0.19	0.143	0.381*	0.238
		30	0.524***	0.429**	0.238	0.143	0.286	0.286	0.238	0	0.143	0.143	0.143	0.286	0.19	0.238	0.143	0.143	0.143	0.238	0.381*	0.286
		20	0.476**	0.429**	0.333	0.238	0.286	0.286	0.238	0.143	0	0.19	0.143	0.286	0.19	0.19	0.143	0.143	0.143	0.286	0.429**	0.286
		10	0.429**	0.333	0.19	0.143	0.238	0.238	0.19	0.143	0.19	0	0.143	0.19	0.19	0.238	0.19	0.143	0.143	0.143	0.476**	0.381*
		5	0.476**	0.381*	0.238	0.143	0.333	0.238	0.19	0.143	0.143	0.143	0	0.19	0.19	0.238	0.143	0.095	0.143	0.19	0.429**	0.333
		1	0.333	0.286	0.19	0.19	0.143	0.19	0.143	0.286	0.286	0.19	0.19	0	0.19	0.381*	0.19	0.238	0.19	0.095	0.429**	0.286
	Median	40	0.381*	0.333	0.238	0.286	0.238	0.238	0.143	0.19	0.19	0.19	0.19	0.19	0	0.286	0.238	0.19	0.19	0.19	0.333	0.238
		30	0.619***	0.524***	0.429**	0.333	0.381*	0.381*	0.381*	0.238	0.19	0.238	0.238	0.381*	0.286	0	0.19	0.238	0.286	0.381*	0.524***	0.429**
		20	0.429**	0.381*	0.286	0.238	0.238	0.238	0.19	0.143	0.143	0.19	0.143	0.19	0.238	0.19	0	0.143	0.143	0.238	0.429**	0.286
		10	0.476**	0.429**	0.238	0.238	0.286	0.238	0.19	0.143	0.143	0.143	0.095	0.238	0.19	0.238	0.143	0	0.143	0.238	0.381*	0.333
		5	0.524***	0.429**	0.238	0.19	0.333	0.286	0.19	0.143	0.143	0.143	0.143	0.19	0.19	0.286	0.143	0.143	0	0.238	0.381*	0.333
		1	0.333	0.238	0.19	0.19	0.143	0.19	0.143	0.238	0.286	0.143	0.19	0.095	0.19	0.381*	0.238	0.238	0.238	0	0.429**	0.286
	Syntax Approach		0.19	0.286	0.524***	0.476**	0.333	0.286	0.381*	0.381*	0.429**	0.476**	0.429**	0.429**	0.333	0.524***	0.429**	0.381*	0.381*	0.429**	0	0.238
	Plain RIC Approach		0.238	0.19	0.429**	0.333	0.19	0.143	0.238	0.286	0.286	0.381*	0.333	0.286	0.238	0.429**	0.286	0.333	0.333	0.286	0.238	0
Sharpe Ratio	Min	40	0	0.238	0.429**	0.429**	0.286	0.286	0.286	0.476**	0.476**	0.381*	0.429**	0.333	0.381*	0.571***	0.429**	0.476**	0.381*	0.333	0.143	0.238
		30	0.238	0	0.333	0.381*	0.19	0.238	0.238	0.476**	0.476**	0.381*	0.381*	0.238	0.333	0.571***	0.429**	0.429**	0.381*	0.19	0.286	0.19
		20	0.429**	0.333	0	0.19	0.238	0.286	0.238	0.333	0.286	0.286	0.238	0.19	0.286	0.429**	0.286	0.238	0.238	0.238	0.524***	0.429**
		10	0.429**	0.381*	0.19	0	0.286	0.286	0.19	0.19	0.238	0.143	0.19	0.238	0.286	0.333	0.19	0.19	0.19	0.19	0.476**	0.286
		5	0.286	0.19	0.238	0.286	0	0.095	0.143	0.286	0.286	0.238	0.286	0.143	0.286	0.429**	0.286	0.333	0.238	0.143	0.286	0.19
		1	0.286	0.238	0.286	0.286	0.095	0	0.143	0.333	0.286	0.286	0.238	0.19	0.286	0.476**	0.333	0.286	0.238	0.19	0.286	0.143
	Mean	40	0.286	0.238	0.238	0.19	0.143	0.143	0	0.238	0.238	0.19	0.19	0.143	0.143	0.381*	0.238	0.238	0.19	0.143	0.333	0.238
		30	0.476**	0.476**	0.333	0.19	0.286	0.333	0.238	0	0.095	0.143	0.143	0.286	0.19	0.19	0.095	0.095	0.143	0.286	0.429**	0.333
		20	0.476**	0.476**	0.286	0.238	0.286	0.286	0.238	0.095	0	0.19	0.143	0.286	0.19	0.19	0.143	0.095	0.143	0.286	0.381*	0.333
		10	0.381*	0.381*	0.286	0.143	0.238	0.286	0.19	0.143	0.19	0	0.19	0.19	0.238	0.238	0.19	0.143	0.143	0.19	0.476**	0.333
		5	0.429**	0.381*	0.238	0.19	0.286	0.238	0.19	0.143	0.143	0.19	0	0.286	0.238	0.238	0.143	0.143	0.095	0.238	0.476**	0.381*
		1	0.333	0.238	0.19	0.238	0.143	0.19	0.143	0.286	0.286	0.19	0.286	0	0.19	0.381*	0.286	0.333	0.238	0.095	0.381*	0.286
	Median	40	0.381*	0.333	0.286	0.286	0.286	0.286	0.143	0.19	0.19	0.238	0.238	0.19	0	0.286	0.19	0.19	0.19	0.238	0.333	0.238
		30	0.571***	0.571***	0.429**	0.333	0.429**	0.476**	0.381*	0.19	0.19	0.238	0.238	0.381*	0.286	0	0.19	0.19	0.286	0.381*	0.524***	0.429**
		20	0.429**	0.429**	0.286	0.19	0.286	0.333	0.238	0.095	0.143	0.19	0.143	0.286	0.19	0.19	0	0.095	0.143	0.238	0.429**	0.286
		10	0.476**	0.429**	0.238	0.19	0.333	0.286	0.238	0.095	0.095	0.143	0.143	0.333	0.19	0.19	0.095	0	0.143	0.286	0.429**	0.381*
		5	0.381*	0.381*	0.238	0.19	0.238	0.238	0.19	0.143	0.143	0.143	0.095	0.238	0.19	0.286	0.143	0.143	0	0.19	0.429**	0.333
		1	0.333	0.19	0.238	0.19	0.143	0.19	0.143	0.286	0.286	0.19	0.238	0.095	0.238	0.381*	0.238	0.286	0.19	0	0.381*	0.286
	Syntax Approach		0.143	0.286	0.524***	0.476**	0.286	0.286	0.333	0.429**	0.381*	0.476**	0.476**	0.381*	0.333	0.524***	0.429**	0.429**	0.429**	0.381*	0	0.238
	Plain RIC Approach		0.238	0.19	0.429**	0.286	0.19	0.143	0.238	0.333	0.333	0.333	0.381*	0.286	0.238	0.429**	0.286	0.381*	0.333	0.286	0.238	0
MDD	Min	40	0	0.19	0.333	0.429**	0.333	0.238	0.095	0.19	0.19	0.286	0.286	0.238	0.19	0.19	0.19	0.238	0.238	0.238	0.143	0.19
		30	0.19	0	0.286	0.381*	0.238	0.238	0.19	0.143	0.19	0.238	0.238	0.238	0.286	0.238	0.143	0.19	0.238	0.238	0.143	0.238
		20	0.333	0.286	0	0.095	0.143	0.143	0.286	0.286	0.238	0.19	0.19	0.143	0.333	0.19	0.143	0.143	0.143	0.143	0.333	0.381*
		10	0.429**	0.381*	0.095	0	0.19	0.19	0.381*	0.381*	0.286	0.238	0.19	0.238	0.381*	0.286	0.238	0.238	0.238	0.238	0.429**	0.429**
		5	0.333	0.238	0.143	0.19	0	0.143	0.286	0.238	0.238	0.286	0.238	0.238	0.333	0.19	0.238	0.238	0.238	0.19	0.286	0.429**
		1	0.238	0.238	0.143	0.19	0.143	0	0.238	0.19	0.143	0.19	0.238	0.143	0.381*	0.143	0.19	0.19	0.19	0.143	0.333	0.286
	Mean	40	0.095	0.19	0.286	0.381*	0.286	0.238	0	0.143	0.238	0.286	0.286	0.238	0.143	0.238	0.19	0.19	0.286	0.238	0.19	0.19
		30	0.19	0.143	0.286	0.381*	0.238	0.19	0.143	0	0.143	0.238	0.238	0.19	0.238	0.19	0.143	0.19	0.19	0.19	0.19	0.19
		20	0.19	0.19	0.238	0.286	0.238	0.143	0.238	0.143	0	0.143	0.143	0.095	0.381*	0.143	0.143	0.19	0.095	0.095	0.238	0.238
		10	0.286	0.238	0.19	0.238	0.286	0.19	0.286	0.238	0.143	0	0.143	0.143	0.333	0.143	0.143	0.143	0.143	0.143	0.333	0.238
		5	0.286	0.238	0.19	0.19	0.238	0.238	0.286	0.238	0.143	0.143	0	0.143	0.429**	0.19	0.19	0.143	0.095	0.143	0.333	0.333
		1	0.238	0.238	0.143	0.238	0.238	0.143	0.238	0.19	0.095	0.143	0.143	0	0.286	0.143	0.095	0.19	0.143	0.048	0.286	0.286
	Median	40	0.19	0.286	0.333	0.381*	0.333	0.381*	0.143	0.238	0.381*	0.333	0.429**	0.286	0	0.381*	0.286	0.333	0.381*	0.286	0.19	0.286
		30	0.19	0.238	0.19	0.286	0.19	0.143	0.238	0.19	0.143	0.143	0.19	0.143	0.381*	0	0.143	0.143	0.143	0.143	0.333	0.286
		20	0.19	0.143	0.143	0.238	0.238	0.19	0.19	0.143	0.143	0.143	0.19	0.095	0.286	0.143	0	0.095	0.143	0.095	0.238	0.286
		10	0.238	0.19	0.143	0.238	0.238	0.19	0.19	0.19	0.19	0.143	0.143	0.19	0.333	0.143	0.095	0	0.143	0.19	0.286	0.333
		5	0.238	0.238	0.143	0.238	0.238	0.19	0.286	0.19	0.095	0.143	0.095	0.143	0.381*	0.143	0.143	0.143	0	0.095	0.333	0.333
		1	0.238	0.238	0.143	0.238	0.19	0.143	0.238	0.19	0.095	0.143	0.143	0.048	0.286	0.143	0.095	0.19	0.095	0	0.286	0.333
	Syntax Approach		0.143	0.143	0.333	0.429**	0.286	0.333	0.19	0.19	0.238	0.333	0.333	0.286	0.19	0.333	0.238	0.286	0.333	0.286	0	0.19
	Plain RIC Approach		0.19	0.238	0.381*	0.429**	0.429**	0.286	0.19	0.19	0.238	0.238	0.333	0.286	0.286	0.286	0.286	0.333	0.333	0.333	0.19	0
Calmar Ratio	Min	40	0	0.19	0.381*	0.429**	0.333	0.333	0.286	0.381*	0.333	0.381*	0.381*	0.333	0.286	0.429**	0.381*	0.333	0.381*	0.333	0.143	0.238
		30	0.19	0	0.333	0.429**	0.19	0.238	0.238	0.333	0.333	0.333	0.333	0.238	0.286	0.429**	0.333	0.333	0.333	0.238	0.286	0.143
		20	0.381*	0.333	0	0.143	0.286	0.286	0.286	0.238	0.286	0.238	0.19	0.19	0.333	0.333	0.286	0.286	0.19	0.238	0.476**	0.429**
		10	0.429**	0.429**	0.143	0	0.286	0.286	0.238	0.19	0.19	0.143	0.19	0.238	0.286	0.286	0.19	0.19	0.19	0.238	0.524***	0.381*
		5	0.333	0.19	0.286	0.286	0	0.143	0.143	0.238	0.238	0.19	0.286	0.19	0.238	0.286	0.238	0.238	0.238	0.19	0.286	0.238
		1	0.333	0.238	0.286	0.286	0.143	0	0.19	0.19	0.238	0.238	0.286	0.238	0.19	0.333	0.238	0.238	0.238	0.238	0.286	0.19
	Mean	40	0.286	0.238	0.286	0.238	0.143	0.19	0	0.19	0.19	0.19	0.238	0.19	0.19	0.286	0.19	0.19	0.238	0.19	0.381*	0.19
		30	0.381*	0.333	0.238	0.19	0.238	0.19	0.19	0	0.095	0.143	0.143	0.19	0.19	0.19	0.095	0.095	0.143	0.238	0.381*	0.333
		20	0.333	0.333	0.286	0.19	0.238	0.238	0.19	0.095	0	0.143	0.143	0.238	0.238	0.19	0.095	0.143	0.143	0.238	0.381*	0.333
		10	0.381*	0.333	0.238	0.143	0.19	0.238	0.19	0.143	0.143	0	0.143	0.143	0.238	0.19	0.143	0.095	0.143	0.143	0.429**	0.286
		5	0.381*	0.333	0.19	0.19	0.286	0.286	0.238	0.143	0.143	0.143	0	0.238	0.286	0.19	0.143	0.143	0.143	0.238	0.429**	0.381*
		1	0.333	0.238	0.19	0.238	0.19	0.238	0.19	0.19	0.238	0.143	0.238	0	0.238	0.286	0.238	0.19	0.19	0.095	0.429**	0.333
	Median	40	0.286	0.286	0.333	0.286	0.238	0.19	0.19	0.19	0.238	0.238	0.286	0.238	0	0.333	0.238	0.238	0.238	0.238	0.333	0.238
		30	0.429**	0.429**	0.333	0.286	0.286	0.333	0.286	0.19	0.19	0.19	0.19	0.286	0.333	0	0.143	0.143	0.19	0.286	0.476**	0.429**
		20	0.381*	0.333	0.286	0.19	0.238	0.238	0.19	0.095	0.095	0.143	0.143	0.238	0.238	0.143	0	0.143	0.19	0.19	0.429**	0.333
		10	0.333	0.333	0.286	0.19	0.238	0.238	0.19	0.095	0.143	0.095	0.143	0.19	0.238	0.143	0.143	0	0.143	0.238	0.429**	0.333
		5	0.381*	0.333	0.19	0.19	0.238	0.238	0.238	0.143	0.143	0.143	0.143	0.19	0.238	0.19	0.19	0.143	0	0.19	0.381*	0.381*
		1	0.333	0.238	0.238	0.238	0.19	0.238	0.19	0.238	0.238	0.143	0.238	0.095	0.238	0.286	0.19	0.238	0.19	0	0.429**	0.333
	Syntax Approach		0.143	0.286	0.476**	0.524***	0.286	0.286	0.381*	0.381*	0.381*	0.429**	0.429**	0.429**	0.333	0.476**	0.429**	0.429**	0.381*	0.429**	0	0.238
	Plain RIC Approach		0.238	0.143	0.429**	0.381*	0.238	0.19	0.19	0.333	0.333	0.286	0.381*	0.333	0.238	0.429**	0.333	0.333	0.381*	0.333	0.238	0

The significance is reported for the following levels: * p<0.1; ** p< 0.05; *** p<0.01.

Figure C.4: Sensitivity of the embedding approach to the similarity measures (i.e. $C_{\Delta t}^{\text{median}}$, $C_{\Delta t}^{\text{min}}$, and $C_{\Delta t}^{\text{mean}}$) (4.2), and the m we use for the similarity threshold $P_{m\,\Delta t}$ between idiosyncratic clauses and market component (4.3). We use $\mathbf{n = 40}$ for the percentile $P_{n\,\Delta t}$ that we use as a sentiment threshold for the daily set of news we trade on $E(n)_{\Delta t}$ (4.5). The results shown are based on 20 subsamples of news, each consisting of 365 randomly selected non-consecutive days. In the x-axis we show the combination of similarity measures and ms tested. For example, "mean_1" indicates that we use the measure $C_{\Delta t}^{\text{mean}}$ and $m = 1$. We benchmark the embedding approach to the syntax and the plain RIC approach.

Table C.4: Kolmogorov-Smirnov two-sample test statistics and significance levels for the idiosyncratic approaches using **n=40** for the percentile $P_{n\,\Delta t}$ that we use as sentiment threshold to define the set of news we will trade on (4.5). The results shown are based on 20 sub-samples of news, each consisting of 365 randomly selected non-consecutive days. For the embedding approach, we compare each measure that we use as similarity score between our set of clauses and the daily *market component* set of news (i.e. $C^{\text{median}}_{\Delta t}$, $C^{\text{min}}_{\Delta t}$, and $C^{\text{mean}}_{\Delta t}$) (4.2), and the m we use for the percentile $P_{m\,\Delta t}$ that we use as similarity threshold for our daily set of idiosyncratic news $I(m)_{\Delta t}$ (4.3). We benchmark the embedding approach to the syntax and the plain RIC approach.

Metric	Measure	m	Min						Mean						Median						Syntax Approach	Plain RIC Approach
			40	30	20	10	5	1	40	30	20	10	5	1	40	30	20	10	5	1		
CAGR	Min	40	0	0.19	0.238	0.286	0.238	0.19	0.333	0.429**	0.286	0.286	0.286	0.286	0.333	0.381*	0.286	0.333	0.286	0.286	0.286	0.286
		30	0.19	0	0.286	0.238	0.19	0.19	0.429**	0.429**	0.333	0.333	0.381*	0.238	0.429**	0.429**	0.333	0.333	0.333	0.238	0.381*	0.238
		20	0.238	0.286	0	0.143	0.19	0.238	0.333	0.333	0.238	0.238	0.19	0.19	0.333	0.333	0.238	0.238	0.19	0.143	0.524***	0.429**
		10	0.286	0.238	0.143	0	0.19	0.238	0.286	0.333	0.238	0.19	0.19	0.143	0.286	0.286	0.19	0.19	0.19	0.143	0.524***	0.429**
		5	0.238	0.19	0.19	0.19	0	0.143	0.381*	0.381*	0.286	0.286	0.238	0.19	0.333	0.381*	0.333	0.286	0.238	0.19	0.524***	0.333
		1	0.19	0.19	0.238	0.238	0.143	0	0.333	0.429**	0.286	0.286	0.286	0.19	0.381*	0.333	0.286	0.286	0.286	0.19	0.476**	0.286
	Mean	40	0.333	0.429**	0.333	0.286	0.381*	0.333	0	0.19	0.238	0.238	0.238	0.333	0.143	0.19	0.19	0.19	0.238	0.238	0.571***	0.476**
		30	0.429**	0.429**	0.333	0.333	0.381*	0.429**	0.19	0	0.19	0.238	0.333	0.333	0.19	0.143	0.19	0.238	0.333	0.333	0.619***	0.571***
		20	0.286	0.333	0.238	0.238	0.286	0.286	0.238	0.19	0	0.143	0.19	0.238	0.238	0.19	0.143	0.143	0.238	0.19	0.524***	0.476**
		10	0.286	0.333	0.238	0.19	0.286	0.286	0.238	0.238	0.143	0	0.095	0.19	0.238	0.19	0.143	0.19	0.143	0.143	0.524***	0.381*
		5	0.286	0.381*	0.19	0.19	0.238	0.286	0.238	0.333	0.19	0.095	0	0.143	0.238	0.238	0.143	0.143	0.143	0.143	0.524***	0.381*
		1	0.286	0.238	0.19	0.143	0.19	0.19	0.333	0.333	0.238	0.19	0.143	0	0.286	0.333	0.238	0.238	0.19	0.143	0.524***	0.333
	Median	40	0.333	0.429**	0.333	0.286	0.333	0.381*	0.143	0.19	0.238	0.238	0.238	0.286	0	0.238	0.19	0.19	0.238	0.286	0.571***	0.476**
		30	0.381*	0.429**	0.333	0.286	0.381*	0.333	0.19	0.143	0.19	0.19	0.238	0.333	0.238	0	0.143	0.143	0.238	0.238	0.619***	0.476**
		20	0.286	0.333	0.238	0.19	0.333	0.286	0.19	0.19	0.143	0.143	0.143	0.238	0.19	0.143	0	0.143	0.19	0.19	0.571***	0.429**
		10	0.333	0.333	0.238	0.19	0.286	0.286	0.19	0.238	0.143	0.19	0.143	0.238	0.19	0.143	0.143	0	0.143	0.19	0.571***	0.381*
		5	0.286	0.333	0.19	0.19	0.238	0.286	0.238	0.333	0.238	0.143	0.143	0.19	0.238	0.238	0.19	0.143	0	0.143	0.524***	0.381*
		1	0.286	0.238	0.143	0.143	0.19	0.19	0.238	0.333	0.19	0.143	0.143	0.143	0.286	0.238	0.19	0.19	0.143	0	0.571***	0.286
	Syntax Approach		0.286	0.381*	0.524***	0.524***	0.524***	0.476**	0.571***	0.619***	0.524***	0.524***	0.524***	0.524***	0.571***	0.619***	0.571***	0.571***	0.524***	0.571***	0	0.333
	Plain RIC Approach		0.286	0.238	0.429**	0.429**	0.333	0.286	0.476**	0.571***	0.476**	0.381*	0.381*	0.333	0.476**	0.476**	0.429**	0.381*	0.381*	0.286	0.333	0
Sharpe Ratio	Min	40	0	0.19	0.238	0.286	0.286	0.238	0.286	0.381*	0.286	0.286	0.286	0.286	0.381*	0.333	0.286	0.333	0.286	0.286	0.286	0.238
		30	0.19	0	0.286	0.286	0.19	0.19	0.429**	0.429**	0.333	0.333	0.333	0.19	0.381*	0.429**	0.381*	0.333	0.286	0.238	0.381*	0.19
		20	0.238	0.286	0	0.143	0.19	0.238	0.286	0.333	0.19	0.238	0.143	0.19	0.286	0.286	0.238	0.19	0.19	0.143	0.524***	0.429**
		10	0.286	0.286	0.143	0	0.19	0.238	0.238	0.286	0.238	0.19	0.143	0.143	0.238	0.238	0.19	0.19	0.143	0.143	0.571***	0.381*
		5	0.286	0.19	0.19	0.19	0	0.19	0.333	0.429**	0.286	0.333	0.238	0.143	0.333	0.333	0.286	0.238	0.19	0.19	0.524***	0.333
		1	0.238	0.19	0.238	0.238	0.19	0	0.381*	0.429**	0.286	0.286	0.238	0.19	0.333	0.333	0.286	0.333	0.286	0.238	0.476**	0.286
	Mean	40	0.286	0.429**	0.286	0.238	0.333	0.381*	0	0.143	0.19	0.143	0.19	0.238	0.238	0.143	0.143	0.19	0.19	0.238	0.571***	0.476**
		30	0.381*	0.429**	0.333	0.286	0.429**	0.429**	0.143	0	0.19	0.238	0.286	0.333	0.238	0.143	0.19	0.238	0.333	0.286	0.619***	0.571***
		20	0.286	0.333	0.19	0.238	0.286	0.286	0.19	0.19	0	0.143	0.19	0.238	0.286	0.143	0.095	0.143	0.238	0.19	0.524***	0.476**
		10	0.286	0.333	0.238	0.19	0.333	0.286	0.143	0.238	0.143	0	0.095	0.238	0.238	0.143	0.143	0.19	0.143	0.19	0.524***	0.381*
		5	0.286	0.333	0.143	0.143	0.238	0.238	0.19	0.286	0.19	0.095	0	0.19	0.238	0.19	0.095	0.19	0.143	0.143	0.524***	0.381*
		1	0.286	0.19	0.19	0.143	0.143	0.19	0.238	0.333	0.238	0.238	0.19	0	0.286	0.286	0.19	0.286	0.19	0.095	0.476**	0.333
	Median	40	0.381*	0.381*	0.286	0.238	0.333	0.333	0.238	0.238	0.286	0.238	0.238	0.286	0	0.238	0.286	0.238	0.238	0.286	0.571***	0.476**
		30	0.333	0.429**	0.286	0.238	0.333	0.333	0.143	0.143	0.143	0.143	0.19	0.286	0.238	0	0.143	0.143	0.19	0.238	0.619***	0.476**
		20	0.286	0.381*	0.238	0.19	0.286	0.286	0.143	0.19	0.095	0.143	0.095	0.19	0.286	0.143	0	0.19	0.143	0.143	0.571***	0.429**
		10	0.333	0.333	0.19	0.19	0.238	0.333	0.19	0.238	0.143	0.19	0.19	0.286	0.238	0.143	0.19	0	0.143	0.238	0.571***	0.429**
		5	0.286	0.286	0.19	0.143	0.19	0.286	0.19	0.333	0.238	0.143	0.143	0.19	0.238	0.19	0.143	0.143	0	0.19	0.524***	0.381*
		1	0.286	0.238	0.143	0.143	0.19	0.238	0.238	0.286	0.19	0.19	0.143	0.095	0.286	0.238	0.143	0.238	0.19	0	0.571***	0.333
	Syntax Approach		0.286	0.381*	0.524***	0.571***	0.524***	0.476**	0.571***	0.619***	0.524***	0.524***	0.524***	0.476**	0.571***	0.619***	0.571***	0.571***	0.524***	0.571***	0	0.333
	Plain RIC Approach		0.238	0.19	0.429**	0.381*	0.333	0.286	0.476**	0.571***	0.476**	0.381*	0.381*	0.333	0.476**	0.476**	0.429**	0.429**	0.381*	0.333	0.333	0
MDD	Min	40	0	0.143	0.238	0.286	0.333	0.238	0.143	0.238	0.238	0.286	0.238	0.238	0.19	0.286	0.238	0.238	0.238	0.238	0.286	0.19
		30	0.143	0	0.238	0.333	0.333	0.238	0.143	0.238	0.238	0.286	0.333	0.238	0.143	0.238	0.19	0.333	0.333	0.238	0.333	0.095
		20	0.238	0.238	0	0.19	0.19	0.19	0.238	0.095	0.143	0.19	0.19	0.19	0.238	0.143	0.143	0.143	0.143	0.143	0.429**	0.286
		10	0.286	0.333	0.19	0	0.095	0.238	0.238	0.19	0.19	0.19	0.143	0.19	0.238	0.238	0.286	0.095	0.143	0.143	0.476**	0.333
		5	0.333	0.333	0.19	0.095	0	0.19	0.238	0.238	0.238	0.143	0.143	0.143	0.286	0.286	0.333	0.095	0.143	0.095	0.524***	0.333
		1	0.238	0.238	0.19	0.238	0.19	0	0.238	0.143	0.19	0.19	0.238	0.143	0.19	0.238	0.238	0.238	0.238	0.143	0.476**	0.238
	Mean	40	0.143	0.143	0.238	0.238	0.238	0.238	0	0.143	0.19	0.238	0.238	0.238	0.095	0.238	0.19	0.238	0.238	0.238	0.333	0.19
		30	0.238	0.238	0.095	0.19	0.238	0.143	0.143	0	0.095	0.143	0.19	0.143	0.19	0.095	0.143	0.143	0.143	0.143	0.381*	0.238
		20	0.238	0.238	0.143	0.19	0.238	0.19	0.19	0.095	0	0.143	0.19	0.19	0.19	0.095	0.143	0.143	0.19	0.143	0.429**	0.286
		10	0.286	0.286	0.19	0.19	0.143	0.19	0.238	0.143	0.143	0	0.095	0.095	0.19	0.19	0.19	0.143	0.143	0.095	0.429**	0.238
		5	0.238	0.333	0.19	0.143	0.143	0.238	0.238	0.19	0.19	0.095	0	0.143	0.238	0.238	0.238	0.095	0.048	0.143	0.524***	0.286
		1	0.238	0.238	0.19	0.19	0.143	0.143	0.238	0.143	0.19	0.095	0.143	0	0.19	0.238	0.238	0.19	0.19	0.095	0.476**	0.333
	Median	40	0.19	0.143	0.238	0.238	0.286	0.19	0.095	0.19	0.19	0.19	0.238	0.19	0	0.238	0.143	0.238	0.238	0.19	0.333	0.19
		30	0.286	0.238	0.143	0.238	0.286	0.238	0.238	0.095	0.095	0.19	0.238	0.238	0.238	0	0.143	0.19	0.238	0.19	0.381*	0.286
		20	0.238	0.19	0.143	0.286	0.333	0.238	0.19	0.143	0.143	0.19	0.238	0.238	0.143	0.143	0	0.238	0.238	0.238	0.333	0.19
		10	0.238	0.333	0.143	0.095	0.095	0.238	0.238	0.143	0.143	0.143	0.095	0.19	0.238	0.19	0.238	0	0.095	0.095	0.476**	0.333
		5	0.238	0.333	0.143	0.143	0.143	0.238	0.238	0.143	0.19	0.143	0.048	0.19	0.238	0.238	0.238	0.095	0	0.19	0.476**	0.333
		1	0.238	0.238	0.143	0.143	0.095	0.143	0.238	0.143	0.143	0.095	0.143	0.095	0.19	0.19	0.238	0.095	0.19	0	0.476**	0.333
	Syntax Approach		0.286	0.333	0.429**	0.476**	0.524***	0.476**	0.333	0.381*	0.429**	0.429**	0.524***	0.476**	0.333	0.381*	0.333	0.476**	0.476**	0.476**	0	0.333
	Plain RIC Approach		0.19	0.095	0.286	0.333	0.333	0.238	0.19	0.238	0.286	0.238	0.286	0.333	0.19	0.286	0.19	0.333	0.333	0.333	0.333	0
Calmar Ratio	Min	40	0	0.19	0.238	0.286	0.286	0.286	0.286	0.333	0.286	0.286	0.286	0.286	0.333	0.333	0.286	0.333	0.286	0.286	0.333	0.286
		30	0.19	0	0.286	0.333	0.238	0.19	0.429**	0.429**	0.333	0.333	0.333	0.238	0.429**	0.381*	0.381*	0.381*	0.333	0.286	0.381*	0.19
		20	0.238	0.286	0	0.19	0.19	0.238	0.238	0.238	0.19	0.19	0.19	0.19	0.286	0.19	0.238	0.238	0.19	0.143	0.524***	0.429**
		10	0.286	0.333	0.19	0	0.238	0.286	0.238	0.238	0.19	0.143	0.095	0.143	0.19	0.19	0.238	0.238	0.19	0.19	0.524***	0.476**
		5	0.286	0.238	0.19	0.238	0	0.19	0.286	0.333	0.238	0.238	0.238	0.143	0.286	0.286	0.238	0.238	0.238	0.143	0.524***	0.333
		1	0.286	0.19	0.238	0.286	0.19	0	0.286	0.381*	0.286	0.238	0.238	0.143	0.286	0.286	0.286	0.286	0.238	0.19	0.524***	0.286
	Mean	40	0.286	0.429**	0.238	0.238	0.286	0.286	0	0.143	0.143	0.19	0.19	0.238	0.143	0.095	0.095	0.143	0.19	0.19	0.571***	0.476**
		30	0.333	0.429**	0.238	0.238	0.333	0.381*	0.143	0	0.238	0.238	0.19	0.286	0.19	0.143	0.19	0.286	0.333	0.286	0.619***	0.571***
		20	0.286	0.333	0.19	0.19	0.238	0.286	0.143	0.238	0	0.143	0.143	0.19	0.19	0.143	0.095	0.19	0.19	0.19	0.524***	0.476**
		10	0.286	0.333	0.19	0.143	0.238	0.238	0.19	0.238	0.143	0	0.095	0.143	0.19	0.19	0.19	0.19	0.143	0.143	0.524***	0.381*
		5	0.286	0.333	0.19	0.095	0.238	0.238	0.19	0.19	0.143	0.095	0	0.143	0.19	0.19	0.19	0.19	0.143	0.143	0.571***	0.429**
		1	0.286	0.238	0.19	0.143	0.143	0.143	0.238	0.286	0.19	0.143	0.143	0	0.19	0.238	0.238	0.238	0.19	0.143	0.524***	0.333
	Median	40	0.333	0.429**	0.286	0.19	0.286	0.286	0.143	0.19	0.19	0.19	0.19	0.19	0	0.143	0.19	0.19	0.19	0.238	0.571***	0.476**
		30	0.333	0.381*	0.19	0.19	0.286	0.286	0.095	0.143	0.143	0.19	0.19	0.238	0.143	0	0.143	0.19	0.19	0.238	0.619***	0.476**
		20	0.286	0.381*	0.238	0.238	0.238	0.286	0.095	0.19	0.095	0.19	0.19	0.238	0.19	0.143	0	0.143	0.143	0.143	0.571***	0.429**
		10	0.333	0.381*	0.238	0.238	0.238	0.286	0.143	0.286	0.19	0.19	0.19	0.238	0.19	0.19	0.143	0	0.143	0.238	0.571***	0.381*
		5	0.286	0.333	0.19	0.19	0.238	0.238	0.19	0.333	0.19	0.143	0.143	0.19	0.19	0.19	0.143	0.143	0	0.19	0.571***	0.381*
		1	0.286	0.286	0.143	0.19	0.143	0.19	0.19	0.286	0.19	0.143	0.143	0.143	0.238	0.238	0.143	0.238	0.19	0	0.524***	0.333
	Syntax Approach		0.333	0.381*	0.524***	0.524***	0.524***	0.524***	0.571***	0.619***	0.524***	0.524***	0.571***	0.524***	0.571***	0.619***	0.571***	0.571***	0.571***	0.524***	0	0.333
	Plain RIC Approach		0.286	0.19	0.429**	0.476**	0.333	0.286	0.476**	0.571***	0.476**	0.381*	0.429**	0.333	0.476**	0.476**	0.429**	0.381*	0.381*	0.333	0.333	0

The significance is reported for the following levels: * p<0.1; ** p< 0.05; *** p<0.01.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**

With my signature I confirm that
- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.