Information Processing and Management 50 (2014) 426–441

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Evaluating and understanding text-based stock price prediction models

Enric Junqué de Fortuny[a,*], Tom De Smedt[b], David Martens[a], Walter Daelemans[b]

[a] Faculty of Applied Economics, University of Antwerp, Prinsstraat 13, B-2000 Antwerp, Belgium
[b] Faculty of Arts, University of Antwerp, Prinsstraat 13, B-2000 Antwerp, Belgium

ARTICLE INFO

Article history:
Received 18 September 2012
Received in revised form 20 August 2013
Accepted 23 December 2013
Available online 31 January 2014

Keywords:
Stock
Prediction
Support Vector Machine
Sentiment mining
Evaluation

ABSTRACT

Despite the fact that both the Efficient Market Hypothesis and Random Walk Theory postulate that it is impossible to predict future stock prices based on currently available information, recent advances in empirical research have been proving the opposite by achieving what seems to be better than random prediction performance. We discuss some of the (dis)advantages of the most widely used performance metrics and conclude that is difficult to assess the external validity of performance using some of these measures. Moreover, there remain many questions as to the real-world applicability of these empirical models. In the first part of this study we design novel stock price prediction models, based on state-of-the-art text-mining techniques to assert whether we can predict the movement of stock prices more accurately by including indicators of irrationality. Along with this, we discuss which metrics are most appropriate for which scenarios in order to evaluate the models. Finally, we discuss how to gain insight into text-mining-based stock price prediction models in order to evaluate, validate and refine the models.

1. Introduction

Is there a way to outperform other investors on the markets? It is a question that has attracted the attention of many a trader since the advent of stock markets in Europe during the late Middle Ages. With the emergence of companies, financial institutions, financial products and government-imposed regulations on these products, the nature of stock markets has changed substantially since those days. Nevertheless, stock price prediction remains an attractive topic for both researchers and investors.[1] During the past decades different theories have been developed to motivate why stock price prediction is not feasible under the assumption of rational actors in a market.

The *efficiënt market hypothesis* was first introduced by Fama (1965) and posits that the financial markets are informationally efficient. This implies that one cannot design a system to predict the change in stock price based on any information because all information is already reflected in the current stock price. Similarly, Malkiel (1985) argues that the stock market prices of assets evolve in a pattern comparable to that of a Random Walk (hence its name *Random Walk Theory*). The implication is that stock prices cannot be predicted better than a "blindfolded chimpanzee throwing darts" at a numerical scale board.

The contributions of this publication are twofold. First, we build empirical models to try to counter-act the validity of the previously mentioned theories, based on the fact that humans do not always act rationally. In order to do so, we combine text-mining techniques in a novel hybrid modelling technique. Second, we discuss the difficulties in evaluating such a model

* Corresponding author. Tel.: +32 32654393.
E-mail address: enric.junquedefortuny@uantwerp.be (E. Junqué de Fortuny).

[1] Querying Google for "stock prediction" reveals over 1.9 million results, including 1360 scientific publications.

0306-4573/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2013.12.002

as a *decision making tool*. We discuss how using many different evaluation metrics can remedy this situation and we show how the model can be used as a *decision support tool* without the aforementioned drawbacks.

2. Empirical research on stock prediction

2.1. Design of empirical models using text mining

A selection of recent empirical research is shown in Table 1, together with the main design choices in these studies. As can be seen from the table, most of the models predicted classes of movement (e.g. up/down) instead of the actual values. Although traditionally most research has centred on various short- and long-term technical performance indicators of a stock (e.g., Lavrenko, Schmill, Lawrie, & Ogilvie (2000)), more recent research has focused on building models based on textual information to perform directional predictions of stock movement (e.g., Schumaker & Chen (2009b); Mittermayer (2006)). The behavior explained by both theories mentioned in the Introduction is based on the assumption that investors act rationally. One way to counter them is to counter the assumption of rationality of the trader. Irrational behavior could for example occur as a reaction to news in the popular written media, i.e. information that comes in the format of text. Since there is a lag between the appearance of an article text and the trading action of the reader, automatic trading systems could outperfom human reaction in a high-frequency trading environment.

2.1.1. Text mining

Text mining concerns the process of automatically extracting novel, non-trivial information from unstructured text documents (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), by combining techniques from data mining, machine learning, natural language processing (NLP), information retrieval (IR) and knowledge management (Mihalcea et al., 2007). Common text mining tasks involve document classification, summarization, clustering of similar documents, concept extraction and sentiment analysis. Text mining has had a wide range of applications to date. Prevalent applications include: forecasting petitions (Suh, Park, & Jeon, 2010), guiding financial investments (Rada, 2008) and sentiment detection (Tang, Tan, & Cheng, 2009; Junqué de Fortuny, De Smedt, Martens, & Daelemans, 2012).

In our setup we look for patterns in the occurrence of words (the so called *bag-of-words*-approach) or the sentiment of the message that have an influence on the stock market price of a commodity. In related work, typically only the direction (rise/fall) of the stock movement is predicted and the patterns come in the form of a linear model, in which each word of a certain vocabulary receives a weight towards the stock price either going up or down. The weighted sum of the word scores of all words in the article is then used in the prediction of a new article. Reported results on independent test sets in terms of accuracy have been in the order of a 10% increase when compared to random predictions (Mittermayer, 2006). We will explore the specifics of text mining for stock price prediction in Section 4 when we discuss the construction of the empirical models.

2.1.2. Combining information

Amongst others, Li, Wang, Dong, and Wang (2011) and Schumaker and Chen (2009b) remark that using only textual information can be too limiting because the approach disregards other (complementary) information. Imagine that a negative news article concerning a certain asset is published during an upward trend of the asset's price. This article might influence the positive trend in a negative way by reducing the slope of that trend, yet the overall trend for the asset can stay upward (see Fig. 1). In this case a negative directional prediction would be wrong, although the impact of the message itself was

Table 1

Literature overview of studies containing text analysis for stock prediction with key design choices. Popular techniques include Naive Bayes (NB), genetic algorithms (GA) and Support Vector Machines (SVM).

Reference	Prediction window	Exchange/index	Technique	Metrics	Target
Wuthrich and Cho (1998)	closing price	Mixed	NB	acc., return	+/±/–
Lavrenko et al. (2000)	1 h	Mixed	NB	profit	++/+/±/–/–
Thomas (2000)	closing price	NASDAQ, NYSE	hybrid GA	excess returns	+/±/–
Gidofalvi (2001)	1 h	NASDAQ	NB	precision/recall	+/±/–
Peramunetilleke (2002)	1-3 h	Currency rates DEM/USD JPY/USD	decision rules	acc.	+/+–/–
Pui Cheong Fung and Xu Yu (2003)	1 h	Hongkong Stock Exchange	SVM	return	+/±/–
Mittermayer (2006)	15 m	S& P 500	SVM	acc., profit, return	+/±/
Zhai et al. (2007)	20 m	BHP Billion Ltd.	SVM	acc., profit	+/–
Schumaker and Chen (2009a)	20 m	S& P 500	SVR	acc., return	+/– and value
Li et al. (2011)	5–30 m	Hang Seng Index	SVM	acc.	+/–
This publications	1 m-64 m/1 day	Euronext Brussels	SVM	acc., AUC, return, Sharpe	+/–

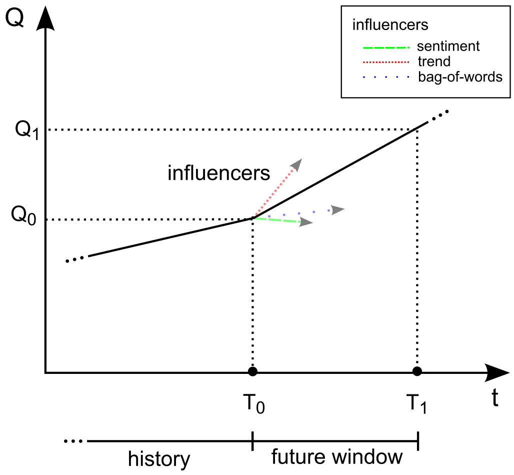

Fig. 1. Multiple influences together determine the actual price as opposed to one variable.

negative. This behavior can be remedied by analysing proxies for the trend of the stock price and incorporating these in the prediction model. Typically, one or more technical indicators are included. In this study we take a similar but novel approach. We analyze three main factors: a series of technical indicators, a bag-of-words score and the sentiment of a news message (which can induce a bias of the investors).

2.2. Proposed model

No clear comparison of the above methods exists, yet each of the authors in Table 1 reports better than random accuracy. We hypothesize that a combination of all of the previous modelling approaches in one hybrid model that contains technical information, text-pattern information and sentiment information should be able to reach even better performance. Our first research question is thus:

Research Question 1 *Can we predict the movement of stock market prices more accurately by including indicators of irrationality next to the traditional trade model features?*

As can be seen in Table 1, prior research has used a range of metrics, with accuracy being the most widely used. We question whether this is actually a good measure based on our own results. This leads us to the second research question:

Research Question 2 *Which metrics are most appropriate for assessing model performance in the context of stock prediction?*

Furthermore, we noticed that many of the models built in the literature use complex modelling techniques that result in 'black-box' models. From the perspective of a trader using the trading tool, however, it is imperative that she can gain insight into these models to assess whether the decision is correct or not.

Research Question 3 *How can we gain insight into text-mining-based stock-prediction models?*

3. Evaluation of empirical models

We have encountered a variety of evaluation metrics, with most studies basing their conclusions solely on one or two metrics (e.g., accuracy). Unfortunately, most of these metrics do not give an accurate representation of the usefulness of the trading model. We argue that one can only gain insights into the validity of a previously constructed model by combining various additional evaluation methods. In this section we consider some of the more frequently used evaluation metrics and their main advantages and drawbacks.

3.1. Discriminatory power

Accuracy measures the percentage of correct predictions out of the total amount of predictions made. More than half of the previous studies we encountered used accuracy to evaluate their models. The problem with using accuracy as a performance measure is that it is difficult to assert whether the built models are valid, due to the fact that the data in the test set is generally not uniformly distributed. For example, consider a trading model that always predicts class 1 (upward price movement). If the target label distribution were skewed so as to contain 70% positive examples (class 1) and 30% negative examples (class 2), the resulting accuracy would be 70%. This is a good result at face value, whereas in reality the model has no predictive power for downward price movement. The accuracy measure implicitly uses a fixed cut-off value (zero) on the output scores of a prediction model to predict whether the output label should be positive or negative. This limitation can also be seen as a limitation on the trading strategy since changing the cutoff value leads to different levels of conservativeness in trading, as such we are only evaluating one possible choice when using accuracy. Many other similar measures that operate on the rows of the confusion matrix, suffer from similar problems (including precision, lift and F scores).[2]

Area Under the Receiver Operating Characteristic Curve (AUC, Fawcett (2006)) is a metric that allows easy evaluation of the discriminative power of a model by measuring the performance of the classification model over the complete range of possible cut-off values. It is a proxy for the probability that the model will classify a randomly chosen positive instance higher than a randomly chosen negative one and as such it is related to the Gini coefficient (Breiman, Friedman, Stone, & Olshen, 1984), i.e. Gini $= 2 \cdot \text{AUC} - 1$. The AUC is a generally accepted performance metric to assess the predictive performance of classification models with respect to ranking in data mining. One of the key advantages of using AUC is its ability to cope with skewed distributions of target label data. Furthermore, it allows for easy comparison with random predictions (i.e., the Random Walk Hypothesis), since a random classifier should result in a AUC value of 50%. Even so, none of the studies we encountered in the literature used this metric.

3.2. Trading simulation

AUC is a useful metric for measuring model discrimination power. Even so, just like accuracy, it proves to be less useful in evaluating the real world operational value of a classifier since it makes assumptions about the data that might not be portable to a real setting. These dynamically changing costs are not easily captured in an aggregate measure. Alternative measures exist that simulate how the model would be used in a real world setting. We will discuss both a proxy for the profit of the trading simulation and the Sharpe ratio.

In our set-up, the only rule is to buy at time t when the model predicts the price is likely to go up within some time frame defined by the lag $t + l$. We sell the stock again at time $t + l$. Given this simple trading strategy, our final revenues and Sharpe ratio can then be evaluated on a test set, simulating a real-life trading scenario.

Profit. Almost every study we found included some form of profit measurement. Some simulate the trading model with a fixed budget (e.g., $50.000) and then report the net profit from a chosen trading strategy. In this set-up, the trading cost is usually assumed to be zero. The problem with this kind of trading assessment is that it is very hard to compare the results to other publications, since many hard-to-distinguish factors can determine the outcome (e.g., starting budget and test set size). A better way to evaluate profit is to use a representation of (excess) return rate. In this study we use the average of the (arithmetic) Rate of Return (ROR) of each trading decision, defined as:

$$ROR = \frac{p_{t+l} - p_t}{p_t}, \tag{1}$$

where p_{t+l} is the selling price of the commodity and p_t the initial buying price.

The average ROR is not without flaws either: particularly it does not take into account the actual risk undertaken by trading upon the built system. In support of the Random Walk Theory, Malkiel (2005) argues that professional investment managers have not been able to consistently outperform their index benchmarks. More specifically, he states that "*no arbitrage opportunities exist that would allow investors to achieve above-average returns without accepting above-average risk*". We should therefore test our models in a trading simulation using a risk-weighted evaluation metric, discussed next.

Sharpe ratio. In finance, risk is often defined in terms of variance of yields (note that this notion of risk is only reasonable under the assumption of an underlying normal distribution). A naturally occurring metric that captures both of these ideas is the Sharpe ratio $S(x)$:

$$S(x) = \frac{r_x - R_x}{\sigma_x}, \tag{2}$$

where x is the investment for the relevant quote symbol, R_x is the risk-free rate of return (a theoretical construct, estimated using real-valued bonds or currency values), r_x is the average return of x using our trading strategy, σ_x the standard deviation of the average return of x. Ideally, we want to generate a value at least higher than zero (profit).

3.3. In-time or out-of-time?

In order to properly evaluate the prediction performance of a model, no training information can be used in the evaluation. Usually a hold-out set is kept aside for evaluation purposes. In time series prediction, an additional concern is that we cannot use any future information in the training phase of a model, this is usually referred to as *out-of-time*. An example that violates both of these principles has received considerable media attention recently. In a publication by Bollen, Mao, and Zeng (2011), a stock prediction model was built, based on Twitter mood prediction. Even though their study has been criticized for containing methodological and representational flaws,[3] a $40 million hedge fund was started based on the

[2] For a full discussion, we refer the reader to Fawcett (2006).

[3] See http://sellthenews.tumblr.com for a full analysis.

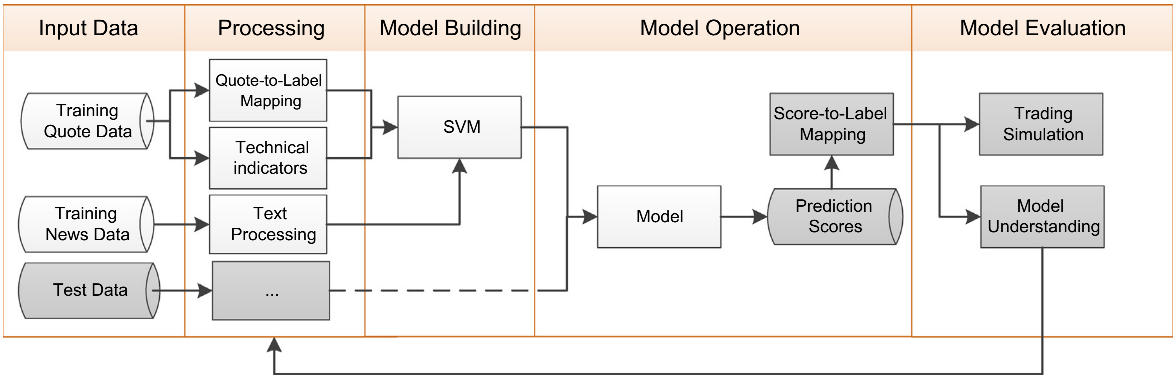

Fig. 2. The main set-up of the model learning, operating and evaluation procedure. As indicated by the back-arrow, models can and should be refined and backtested in an iterated manner during model operation.

technique, receiving a nomination for the 9th annual Awards for Excellence in Trading and Technology Europe 2011 for the most innovative trading Firm.[4]

4. Material and methodology

The main set-up of the training and testing of our stock prediction model is displayed in Fig. 2. In a first phase both stock tick data and stock news data are gathered and processed to two types of features: technical indicators and text-related features. These are given as input to a Support Vector Machine (SVM) learner after which an "optimal" model is generated. In order to evaluate the model, its output scores are converted by a score-to-label mapping and compared to the future quote price evolution. In the next sections we will briefly discuss each of these components.

4.1. Support Vector Machines

We chose the Support Vector Machine as the main driver for model generation since it is well established to be successful for text mining (Cohen & Hersh, 2005; Tang et al., 2009) and more specifically for stock prediction (as can be seen from Table 1). The SVM is a learning procedure based on the statistical learning theory (Vapnik, 1995). Given a training set of m-dimensional input vectors $\mathbf{x} = \{x_1, \ldots, x_m\}$ and corresponding binary class labels $y_i \in \{-1, +1\}$, the SVM classifier constructs a hyperplane in a feature space, induced by the non-linear function φ. According to Vapnik's original formulation, the classifier should satisfy the following conditions:

$$\begin{cases} \mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i) + b \geqslant +1, & \text{if } y_i = +1 \\ \mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i) + b \leqslant -1, & \text{if } y_i = -1 \end{cases}$$

which is equivalent to (for a dataset of size n)

$$y_i(\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i) + b) \geqslant 1, \quad i = 1, \ldots, n. \tag{3}$$

The non-linear function $\boldsymbol{\varphi}(\cdot)$ maps the input space to a high (possibly infinite) dimensional feature space. In this feature space, the above inequalities construct a hyperplane $\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}) + b = 0$, that discriminates between the two classes. By minimizing $\mathbf{w}^T\mathbf{w}$, the margin between both classes is maximized.

In primal weight space the classifier takes the form

$$y(\mathbf{x}) = \operatorname{sign}(\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}) + b),$$

however, it is never evaluated in this form. The exact details of the dual form of SVMs are beyond the scope of this report, but in summary, the problem is recast to an optimization problem that is solved using Lagrange multipliers, leading to the following classifier:

$$y(\mathbf{x}) = \operatorname{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right), \tag{4}$$

where $K(\mathbf{x}_i, \mathbf{x}) = \boldsymbol{\varphi}(\mathbf{x}_i)^T\boldsymbol{\varphi}(\mathbf{x})$ is a positive definite kernel satisfying the Mercer theorem conditions. No explicit construction of the nonlinear mapping $\boldsymbol{\varphi}(\mathbf{x})$ is needed, we only need to choose a kernel function K. For the kernel function $K(\cdot,\cdot)$, one typically has the following choices:

[4] http://www.derwentcapitalmarkets.com/.

$$K(\mathbf{x},\mathbf{x}_i) = \mathbf{x}_i^T\mathbf{x} \quad \text{(linear kernel)}$$
$$K(\mathbf{x},\mathbf{x}_i) = \left(1+\frac{\mathbf{x}_i^T\mathbf{x}}{c}\right)^d \quad \text{(polynomial kernel)}$$
$$K(\mathbf{x},\mathbf{x}_i) = e^{\frac{-\|\mathbf{x}-\mathbf{x}_i\|_2^2}{\sigma^2}} \quad \text{(RBF kernel)}$$

where $d,\ c,\ \sigma$ and θ are constants. Throughout this study we will be using a linear kernel due to the high dimensionality of the data and speed considerations.

4.2. Input data

4.2.1. Document data acquisition and selection

The corpus used in this study comprises all articles published in on-line versions of all major Flemish newspapers in 2007 until March 2012. This leads to a corpus of over 671.751 articles, each of them timestamped on a minute level granularity. An overview of all of the newspapers included in this study is displayed in Table 2.

All articles were gathered using a custom built web-crawler. The crawler extracts articles from the sources' websites using their built-in search functionalities (after obtaining the permission to do so). The crawling process is the equivalent of a typical database selection process in which relevant data are selected using the given query criteria. In this case, the query keywords were the names of the organization behind the stock symbols. An overview of all stock symbols and their search query keywords ("search keys") is displayed in Table 3.

After the filtering process, the scope of the textual data is reduced to include only that part of the article that is relevant to the stock symbol. The following cases are considered:

1. Include only the headline.
2. Use all textual data of the article.
3. Use only textual data in the same paragraph as the first occurrence of the key word: the paragraph is defined as one sentence before, the containing sentence and one sentence after the relevant sentence.

The rationale behind the third approach is that an article can contain many stock symbol names at the same time or switch tone. For each of these cases we consider both the textual input, as well as the sentiment polarity score after sentiment analysis.

Table 2
News sources include in this study and their amount of readers.

Source	#Readers[a]
De Redactie	146,250
De Morgen	256,800
GVA	395,700
HBVL	423,700
Nieuwsblad	1,002,200
De Standaard	314,000
De Tijd	123,300
HLN	1,125,600

[a] Counted by the amount of readers of the printed version except for "De Redactie" which does not exist in a printed format. Instead, we used the number of unique visitors per day in 2009 as an estimation. *Source*: belga/odbs.

Table 3
Overview of stock symbols (Euronext Brussels Stock Exchange) and the search query keyword used to find related news.

Symbol	Full name	Search key
AGS	Ageas	ageas
BELG	Belgacom	belgacom
DELB	Delhaize Group	delhaize
GBLB	Groupe Bruxelles Lambert SA	gbl
ABI	Anheuser-Busch InBev NV	inbev
KBC	KBC Groep NV	kbc
MOBB	Mobistar SA	mobistar
NYR	Nyrstar NV	nyrstar
TNET	Telenet Group Holding NV	telenet
UCB	UCB SA	ucb
UMI	Umicore SA	umicore

4.2.2. Document data preprocessing

In a first step, every article in the corpus has to be lemmatized in order to reduce all of the words in the corpus to their canonical form. Afterwards, all known stop-words are deleted from the corpus. Applying these two steps lowers variability of concept expression in the corpus and allows the learner to focus on content words.

Given the clean corpus D, we build a dictionary containing all of the m words contained in the corpus. With this dictionary, each of the individual documents d can now be represented as a bag-of-words vector $[w_1 w_2 \ldots w_m]$. Aggregating all of these row-vectors, leads to a high-dimensional and very sparse matrix in which each element $w_{i,j}$ contains the amount of occurrences of word i in document j. In order to be able to compare documents of different lengths each row (document) of this matrix is normalized on the total amount of words, leading to a matrix with term-frequencies (tf). This matrix is then rescaled by the inverse document frequency (idf), leading to the input matrix tfidf:

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \times \mathrm{idf}(t,D) \tag{5}$$

$$\mathrm{idf}(t,d,D) = \log \frac{|D|}{1 + |\{d \in D | t \in d\}|}, \tag{6}$$

where $|D|$ is the cardinality for the set of documents D. This helps to prevent commonly occurring words to dominate the learning procedure, since these do not contain discriminatory information. Empirical research has shown this approach to be a valuable transformation (consult Sebastiani (2002) for more information). Note that this scaling is of course based on information from the training set only. Each row in the resulting matrix represents one of the input vectors $\mathbf{x}_i$ that are given as input to the SVM.

4.2.3. Document sentiment polarity

For sentiment analysis, we used the previously created Pattern module for Python (De Smedt & Daelemans, 2012a). The module contains a suite of tools for web mining and text mining, including a subjectivity lexicon of over 3000 Dutch adjectives that occur frequently in product reviews, manually annotated with scores for polarity (positive or negative between +1.0 and −1.0) and subjectivity (objective or subjective between +0.0 and +1.0). For example: 'boeiend' (fascinating) has a positive polarity of +0.9 and 'belabberd' (lousy) has a negative polarity of −0.6. A similar approach with one axis for polarity and one for subjectivity is used by Esuli and Sebastiani (2006) for English words. The Pattern module includes an algorithm that further refines the score for each adjective by looking at preceding adverbs (e.g., extremely fascinating), preceding negation words (e.g., not fascinating) and subsequent exclamation marks.

In previous research, the lexicon was tested with a set of 2000 Dutch online book reviews. Each review also has a user-given star rating. The test set was evenly distributed over negative opinion (star rating 1 and 2) and positive opinion (star rating 4 and 5). The average score of adjectives in each review was then compared to the original star rating, with a precision of 72% and a recall of 82% (De Smedt & Daelemans, 2012b).

In our approach, we calculate the polarity of each adjective that occurs in the input text. The aforementioned third method discussed in Section 2.1.1, that is, using only textual data in the same paragraph as the first occurrence of the key word, is expected to yield a more reliable correlation between the entity being mentioned (the 'target' of the sentiment) and the adjective's polarity score, contrary to measuring all adjectives in the article. A similar approach for target identification with a 10-word window is used by Balahur et al. (2010). They report improved accuracy when compared to measuring all words in the article. This results in a set of 274,014 assessments, where one assessment corresponds to the sentiment score linked to a stock symbol at a particular time. For example, the following assessment (translated from Dutch to English) scores −0.17:

> ''Belgium's largest electricity producer threatens to suspend investments if the nuclear tax is raised. 'If the Belgian State would not honor its commitments, GDF Suez would be forced to revise its policy in terms of investment, employment, education and patronage entirely.' GDF Suez-Electrabel, by far the largest electricity producer in Belgium, fires a booming shot across the bow of the federal governement negotiators.''

4.2.4. Technical indicators

We considered four popular technical indicators in our approach, all of which are based on a series of price ticks $P = \{p_1, p_2, \ldots, p_n\}$ leading up to the last known price p_n (Bodie, Kane, & Marcus, 2008). The length of the series was chosen to be $n = 5$ ticks, which for the one-day-ahead setting corresponds to a period of a week. For the minutes-ahead setting, we kept the amount of ticks, but sampled at higher frequency as well (30 min and 5 min). The exact values of these windows do not matter, but it is important that the technical indicators should catch the trends on the same level of granularity as the prediction window size.

Relative Strength Index (RSI) is an indicator of the historical strength or weakness of a stock over a series of price ticks P is defined as:

$$RSI(P) = 100 - \frac{100}{1 + RS(P)} \tag{7}$$

$$RS(P) = \frac{\text{average gain}}{\text{average loss}} \tag{8}$$

According to Wilder (1978), the creator of this technical indicator, a stock price should be considered overbought when the price moves up very rapidly ($RSI > 70$). Likewise, when the price falls very rapidly (typically $RSI < 30$) it should be considered oversold.

Williams %R is an oscillator index relating the current price p_n to the highest and lowest price of the series P.

$$R(P) = \frac{p_n - \min_{p_i \in P} p_i}{\max_{p_i \in P} p_i - \min_{p_i \in P} p_i}, \tag{9}$$

Psychological Line (PSY) is the technical variant of a sentiment indication and is defined as the percentage of the number of rising periods over the total number of periods considered in the series P:

$$PSY(P) = 100 \times \frac{|\{p_i | p_i \in P \wedge p_i > p_{i-1}\}|}{|P|}. \tag{10}$$

The Bias indicator assesses the behavior of the market in the given period P as "bullish", "bearish" or neutral. Once identified, a trading strategy is recommended to counter-act the market. We did not explicitly code this behavior in our trading strategy, but include it in the features since its information is relevant to the movement of the stock. The bias is defined as:

$$BIAS(P) = 100 \times \frac{p_n - m(P)}{m(P)} \tag{11}$$

$$m(P) = \frac{1}{|P|} \sum_{p_i \in P} p_i \tag{12}$$

Combining Eqs. (7)–(12) for each document based on the document's appearance time leads to a final input vector $\mathbf{x}_i$ that can be given as an input to the SVM.

4.3. Target data

4.3.1. Stock tick data acquisition

We have gathered two datasets containing quote data for all of the stock symbols in Table 3 on the Euronext Brussels Stock Exchange. The first dataset contains quote ticker data on a low-granularity level (per-day) over a period of three years, from January 1, 2007 to March 25, 2012. On a high-granularity level (per-minute), we have gathered tick data for the same stock symbols for a period of four months from January 1, 2012 up to March 5, 2012.

4.3.2. Quote-to-label mapping

In order to simplify the learning system, the problem is reduced to that of predicting a subset of possible stock movements, aggregated in classes (e.g., increase (+), stable (±), decrease (−), see Table 1). We chose binary classification (up/down movement) or the *directionality* of the movement of a stock quote as opposed to predicting the true value. The target prediction labels are based on the relative movement Δ_r of the stock quote as compared to the last-known quote tick data. The relative movement of a quote is defined as:

$$\Delta_r = \frac{Q_1 - Q_0}{Q_1} \tag{13}$$

In our two-class setting, we defined a class for a positive inclination (i.e., $\Delta_r \geqslant 0$) and one for a negative inclination ($\Delta_r < 0$).

For the one-day-ahead setting, the labels are based on opening and closing quotes. For the minutes-ahead setting, we experimented with different time windows after the appearance of an article (on a logarithmic scale, with a lag l ranging from 2^0 to 2^6 min). This time window was chosen in accordance with results from previous empirical experiments from the literature (Table 1). Note that there is no high-granularity level information available outside of office hours of the stock exchange, thus any data gathered during this period was dropped from the dataset (Fig. 3). Previous studies have tried interpolating the values, but we argue that this is a very rough approximation at best when working on a minute granularity level.

4.4. Evaluation

In order to ensure good generalization properties, the SVM uses separate datasets for training, validation and testing (as displayed in Fig. 4). The test set is withheld from the training process and is only used in the evaluation of the trained model. Using a 5-fold in-time cross validation scheme, the kernel and regularization parameters for model construction are determined by training a model with this parameter on a subset of the data (fold training set, displayed in light gray), and then evaluating it on another part (the validation set, displayed in dark gray). By repeating this process over all five folds, we

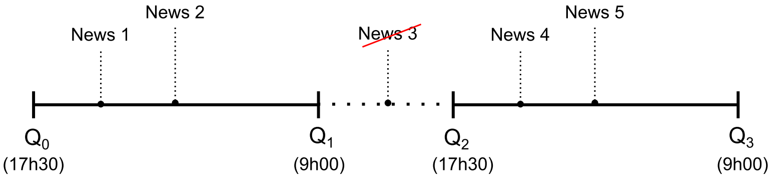

Fig. 3. Time window and relative quote definition.

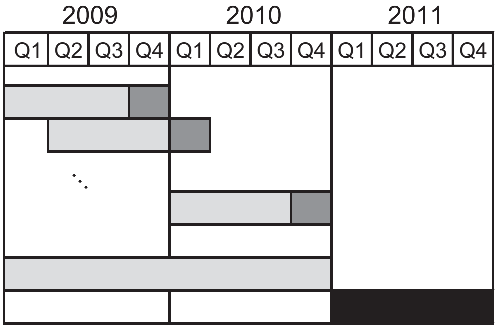

Fig. 4. Out-of-time train/test data split for the one-day-ahead setting.

ensure a more robust evaluation of the model. Once the optimal parameters have been determined, a final model can be built on the full training set. Note that no out-of-time information may be used when building the final system to avoid using information of a future event in a prediction since this information would not be available in a trading system either.

In order to illustrate the problems with some evaluation metrics as mentioned in Section 3, we will include four evaluation metrics in this study: accuracy, AUC, return and Sharpe ratio.

5. Results

5.1. Individual model training

In a first batch of experiments we built 56 models, one for each of the types of features on the seven different minute lags. For both the bag-of-words and the sentiment we considered three different partitions of the texts (as previously explained in Section 4.2.1).

5.1.1. Prediction accuracy

The accuracy of each of the models is displayed in Table 4 for the bag-of-words (BOW), sentiment and technical indicator (TI) approach.[5] Additionally, the rank of the performance over all lags is displayed. As can be observed in Table 4, there is no clear winning design choice when looking at the accuracies. As discussed before, it is difficult to assess from these numbers which model performed better, because of the tacit threshold implied by the accuracy metric. A score greater than zero will be classified as positive news, whereas a score lower than zero will be classified as negative news.

The problem is that it is very difficult to assess how these models would work within a trader's profile. This is illustrated in Fig. 5, where we plotted the accuracy for the trading-behavior spectrum from very conservative (left) to very aggressive (right). Let us consider a more conservative trader that only wants to trade when the model is relatively sure about its decisions. This particular trader might want to trade only at a certain score percentile of the possible trades (indicated by a red mark). The accuracies for this instance of the model threshold choice are clearly better than those of the zero threshold for most models, something that was not immediately obvious from Table 4. The exact differences for both trading profiles are displayed in Table 5. One explanation for this much better behavior is that the input (and by extension the output) of the model is quite noisy and there is no distinction between articles with a relatively low positive score or very high positive score since both get labelled as 'positive news'.

5.1.2. Area under curve

AUC implicitly aggregates all possible trading behaviors in one metric by averaging over all trading thresholds. The AUC values of all models over the test set are displayed in Table 6. Using AUC it is much easier to compare our models to a random

[5] Note that all the accuracy and AUC numbers reported in this study are rounded up to two digits after the decimal separator. We did not deviate from this convention for comparison and representation purposes. This does, of course, not mean that all these digits should be considered significant since this depends on the size of the actual test set.

Table 4
Prediction ability (accuracy) of the trained models on the test set and the average ranking of each of the options (columns).

Lag *t*	Bag-of-words			Sentiment			TI
	Full text	Paragraph	Title	Full text	Paragraph	Title	
1 m	50.00%	46.13%	54.78%	57.37%	43.77%	**69.05**%	47.19%
2 m	58.43%	44.35%	55.06%	**61.22**%	60.38	29.76	59.83%
4 m	53.65%	51.19%	43.26%	**57.69**%	56.55	34.52	57.58%
8 m	53.09%	53.27%	46.63%	54.81	46.01%	44.05	**55.34**%
16 m	52.81%	53.87%	49.72%	50.96	49.84%	40.48	**55.06**%
32 m	46.63%	50.60%	51.97%	50.32	49.20%	**63.10**	46.07%
64 m	45.51%	**53.57**%	50.28%	45.83%	46.01%	50.00%	46.91%
24 h	50.33%	49.26%	50.52%	49.12%	49.22%	48.78%	**51.10**%
Rank	4.4	3.8	3.9	3.3	4.8	5.0	**3.0**

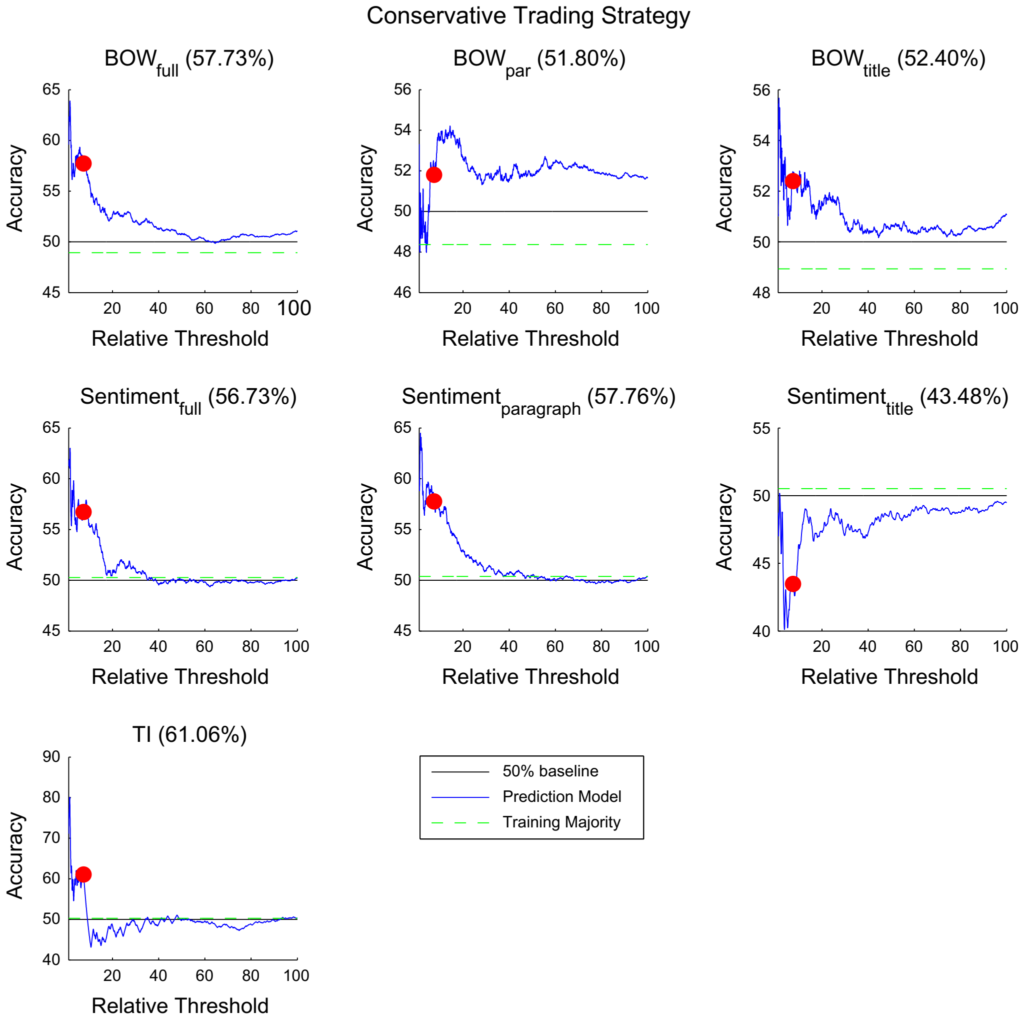

Fig. 5. Accuracy result for varying thresholds in the one-day-ahead prediction setting. A lower relative threshold entails trading only when a large increase in stock price is predicted and thus corresponds to a more conservative trading strategy. Likewise, a higher relative threshold corresponds to a more aggressive trading strategy. The more conservative threshold (marked by a red circle at a threshold value of 7.50%) clearly outperforms a more aggressive trading strategy (e.g. choosing a threshold of 50%).

prediction model since a random prediction model has an AUC of exactly 50%. From Table 6 we can discern that, according to AUC, it is not always possible to build a model that performs better than random, but we can highlight some interesting results that stand out. The technical indicators perform very well. This suggests that they are indeed useful for predicting stock price behavior. This comes as no surprise since technical indicators have been used in trading tools in the past. Technical indicators perform significantly better than all other models (including a completely random model) except for the sentiment-title model (using a Wilcoxon signed rank test, significance level $p < 0.05$).

Table 5
Accuracy results for different trading strategies in the one-day-ahead setting.

Trading strategy	Bag-of-words (%)			Sentiment (%)			TI (%)
	Full text	Paragraph	Title	Full text	Paragraph	Title	
Conservative	57.73	51.80	52.40	56.73	57.76	43.48	61.06
Aggressive	51.06	51.64	51.06	50.27	50.38	49.50	50.27

Table 6
Discrimination ability (AUC) of the trained models on the test set.

Lag *t*	Bag-of-words			Sentiment			TI
	Full text	Paragraph	Title	Full text	Paragraph	Title	
1 m	48.70%	48.69%	**55.18%**	50.07%	49.77%	53.81%	52.87%
2 m	49.03%	47.33%	51.19%	48.99%	49.26%	43.08%	**56.03%**
4 m	53.30%	50.19%	40.93%	47.62%	47.67%	39.25%	**55.35%**
8 m	56.82%	55.16%	47.12%	48.63%	51.72%	43.44%	**58.22%**
16 m	54.28%	54.78%	46.75%	48.66%	52.11%	49.35%	**54.91%**
32 m	51.36%	49.11%	50.99%	46.36%	45.66%	**57.33%**	51.27%
64 m	45.90%	52.60%	49.25%	44.98%	43.64%	39.12%	**53.20%**
24 h	50.29%	50.73%	50.87%	50.26%	50.13%	48.33%	**52.16%**
Rank	3.4	3.9	3.9	5.1	4.9	5.4	**1.5**

The sentiment results are inconsistent and often underperform the other models as compared to the bag-of-words approach or the technical indicators (often even underperforming a random classifier). One possible explanation is that sentiment from one context is not portable to a different one. We hypothesize that, in contrast to the domains on which the sentiment extraction model was trained (book reviews), sentiment in financial texts can be found more predominantly in nouns and verbs rather than in adjectives, on which the current model is based. Another explanation is that the sentiment by itself does not provide enough information and needs to be used conjointly with other information. Although there is strong evidence to believe the sentiment model simply underperformed, it remains difficult to trace the exact reasons and magnitude of the low discriminative power of sentiment. This is one of the drawbacks of using an aggregate measure such as sentiments, as we will see, the BOW approach does not suffer from this flaw.

5.1.3. Rate of return

Table 7 shows the returns from buying/reselling one share of the stock symbols of which the model thinks it is likely to go up in our simple trading simulation. Note that N.A. means that the model did not decide to buy any of the stock symbols based on the articles in the test set. This behavior is an indication of possible bad convergence during the learning procedure since it has a strong bias to classify news as being negative. This is known to happen to SVMs after bad convergence of the internal Sequential Minimal Optimization algorithm. (See Table 8).

Interestingly though, all short trading time span systems ($\leqslant 8$ m) performed positively in terms of return. This corresponds to our original hypothesis: the behavior that we attempt to capture using our models occurs in this time period since the original premise is that we want to predict how traders react to a news article. Given a conscious and informed trader, it is reasonable to assume that the bulk of these actions should be visible in the first minutes after the appearance of the news article. But we must be careful to watch out for data dredging [6] when making any ex post conclusions about the exact lag at which the largest effect of the model can be observed.

As mentioned before (Section 3), an important drawback of using returns is that it is difficult to compare the resulting values, since not all of the generated models in our experimental study have the same test set size, furthermore, these values do not take into account the risk undertaken by the trader when performing such actions.

5.1.4. Sharpe ratio

The risk-free rate was chosen to be 0.07%, based on the average of the AAA-rated Euro area central government bonds' yield rates for March 2012. The resulting Sharpe ratios for our trading simulation are displayed in Table 6, where N.A. again indicates that the model did not decide to predict a positive directionality on any of the news articles published in the test month. The Sharpe ratio, by definition, is closely related to the return rate and therefore, similar effects can be observed. The actual values are somewhat smaller due to the risk factor being taken into account. The risk weighing effect is more noticeable in some models than others (e.g., the 1 m high frequency setting): although return rates were very high, the Sharpe ratios are not as good. This indicates that the trading strategy for the one-day-ahead settings are not without risk, something we could not have known by only looking at the return rates.

[6] Data dredging is the inappropriate use of data mining to reveal misleading relationships in data.

Table 7
Average monthly return rates on the test set in percentage (buy all positives/maximum buy selected).

Lag t	Bag-of-words			Sentiment			TI
	Full text	Paragraph	Title	Full text	Paragraph	Title	
1 m	0.185×10^{-3}	$\mathbf{0.205 \times 10^{-3}}$	0.189×10^{-3}	N.A	0.195×10^{-3}	N.A	0.190×10^{-3}
2 m	0.273×10^{-3}	0.287×10^{-3}	0.283×10^{-3}	N.A	N.A	$\mathbf{0.496 \times 10^{-3}}$	0.284×10^{-3}
4 m	0.268×10^{-3}	0.283×10^{-3}	0.266×10^{-3}	N.A	N.A	$\mathbf{0.692 \times 10^{-3}}$	0.269×10^{-3}
8 m	0.098×10^{-3}	0.105×10^{-3}	0.119×10^{-3}	N.A	$\mathbf{0.151 \times 10^{-3}}$	N.A	0.105×10^{-3}
16 m	-0.053×10^{-3}	-0.034×10^{-3}	-0.066×10^{-3}	N.A	N.A	$\mathbf{0.432 \times 10^{-3}}$	-0.053×10^{-3}
32 m	-0.056×10^{-3}	-0.012×10^{-3}	-0.047×10^{-3}	N.A	N.A	N.A	-0.060×10^{-3}
64 m	-0.158×10^{-3}	-0.108×10^{-3}	-0.122×10^{-3}	N.A	N.A	N.A	-0.125×10^{-3}
24 h	0.277×10^{-3}	$\mathbf{0.321 \times 10^{-3}}$	0.278×10^{-3}	0.156×10^{-3}	0.158×10^{-3}	0.231×10^{-3}	0.165×10^{-3}
Rank	5.1	**2.6**	4.3	4.4	3.8	3.4	4.5

Table 8
Average Sharpe ratio of the AUC-trained models on the test set (buy all positives/maximum buy selected).

Lag t	Bag-of-words			Sentiment			TI
	Full text	Paragraph	Title	Full text	Paragraph	Title	
1 m	-1.484×10^{-3}	-0.852×10^{-3}	-1.354×10^{-3}	N.A.	-1.140×10^{-3}	N.A.	-1.344×10^{-3}
2 m	0.954×10^{-3}	1.242×10^{-3}	1.182×10^{-3}	N.A.	N.A.	$\mathbf{4.646 \times 10^{-3}}$	1.193×10^{-3}
4 m	0.711×10^{-3}	1.003×10^{-3}	0.676×10^{-3}	N.A.	N.A.	$\mathbf{8.008 \times 10^{-3}}$	0.730×10^{-3}
8 m	-2.202×10^{-3}	-2.022×10^{-3}	-1.863×10^{-3}	N.A.	-1.351×10^{-3}	N.A.	-2.087×10^{-3}
16 m	-3.658×10^{-3}	-3.323×10^{-3}	-3.817×10^{-3}	N.A.	N.A.	$\mathbf{2.645 \times 10^{-3}}$	-3.657×10^{-3}
32 m	-2.305×10^{-3}	-1.899×10^{-3}	-2.229×10^{-3}	N.A.	N.A.	N.A.	-2.342×10^{-3}
64 m	-2.464×10^{-3}	-2.081×10^{-3}	-2.220×10^{-3}	N.A.	N.A.	N.A.	-2.240×10^{-3}
24 h	0.293×10^{-3}	$\mathbf{0.344 \times 10^{-3}}$	0.293×10^{-3}	0.157×10^{-3}	0.159×10^{-3}	0.257×10^{-3}	0.168×10^{-3}
Rank	5.6	3.1	4.8	3.1	4.3	**2.1**	5.0

5.2. A hybrid model

In Section 2.1.1 we stated that all the previously mentioned input variables create a joint effect on the movement of the stock price (i.e., there is some complementarity between the effect of each of the individual variables). The fact that the different models react differently in terms of AUC and Sharpe ratio adds supports to this claim. In this section we build a hybrid system that includes all three of the previous types of input variables. That is, for each of the text selection cases (full text, paragraph and title only), we combine the information from the bag-of-words, the sentiment and the technical indicator approach.

The (highly dimensional) bag-of-words is not maintained in its full form, but only the scores of the bag-of-words models are retained. These are then combined with the sentiment score and the technical indicators. To include the interaction effect between the various components of the system, we included interaction variables for each possible combination of technical indicators with either sentiment or bag of words. This leads to a surplus of 90 variables[7] for the minutes-ahead setting and 30 variables for the one-day-ahead setting. The resulting input vector is thus of the following form:

$$\mathbf{x}_i = [\mathrm{TI}_1 \cdots \mathrm{TI}_{30} \quad \underbrace{O_{title} O_{paragraph} O_{full}}_{\text{BOW SVM scores}} \quad \underbrace{S_{full} S_{par} S_{title}}_{\text{Sentiment scores}}]$$

The results of our hybrid experiments are displayed in Table 9.

We encountered similar ambiguous results depending on the metric used for the hybrid model. In terms of the average Sharpe ratio and returns, the hybrid model roughly shows the same behavior: the faster acting models outperform the slower models with the exception of the one-day-ahead model. In terms of return rates and Sharpe ratio the models reach some high levels, but this is countered by the fact that some of the built models compare equal or slightly worse than the individual models.

In conclusion, let us return to our original research question: *can we predict the movement of stock market prices more accurately by including indicators of irrationality next to traditional trade model features? (RQ2)* Based on the results presented in this section, we can only accept the nuanced answer that for certain trading strategies (one example being the conservative trading strategy highlighted before), it is indeed possible to predict stock price movement better using textual information.

[7] 5 technical indicators × 3 time windows (week, half hour, 5 min) × 3 document contexts (full, paragraph, title), and this for both the sentiment as well as the bag-of-words approach.

Table 9
Test results for the hybrid models on different time lag settings.

Lag *t*	Accuracy			AUC		
	Full text	Paragraph	Title	Full text	Paragraph	Title
1 m	55.38%	55.23%	**65.38%**	**51.53%**	48.27%	44.61%
2 m	60.22%	**60.47%**	59.62%	**52.69%**	49.32%	44.44%
4 m	**61.29%**	55.23%	59.62%	**56.35%**	39.23%	39.17%
8 m	**51.61%**	43.60%	46.15%	48.20%	55.37%	**57.93%**
16 m	46.77%	50.58%	**57.69%**	52.23%	51.02%	**56.97%**
32 m	45.16%	44.19%	**46.15%**	**54.62%**	52.58%	53.07%
64 m	47.85%	**59.30%**	53.85%	48.78%	43.41%	**60.71%**
24 h	49.44%	50.28%	**55.11%**	45.23%	**51.54%**	50.04%
Rank	1.8	2.3	2.0	2.1	2.3	1.6
	Rate of return			Sharpe ratio		
1 m	0.083×10^{-3}	**0.131×10^{-3}**	0.099×10^{-3}	-4.735×10^{-3}	**-3.353×10^{-3}**	-4.449×10^{-3}
2 m	N.A	N.A	**0.160×10^{-3}**	N.A	N.A	**-2.201×10^{-3}**
4 m	**0.267×10^{-3}**	N.A	0.075×10^{-3}	**0.716×10^{-3}**	N.A	-3.702×10^{-3}
8 m	0.141×10^{-3}	**0.212×10^{-3}**	0.127×10^{-3}	-1.693×10^{-3}	**-0.401×10^{-3}**	-2.079×10^{-3}
16 m	-0.123×10^{-3}	**-0.021×10^{-3}**	0.025×10^{-3}	-5.329×10^{-3}	-3.792×10^{-3}	**-3.401×10^{-3}**
32 m	-0.093×10^{-3}	-0.035×10^{-3}	**-0.025×10^{-3}**	-3.519×10^{-3}	-2.959×10^{-3}	**-2.517×10^{-3}**
64 m	-0.675×10^{-3}	-0.578×10^{-3}	**-0.254×10^{-3}**	-6.267×10^{-3}	-5.746×10^{-3}	**-3.554×10^{-3}**
24 h	**0.282×10^{-3}**	0.252×10^{-3}	0.056×10^{-3}	**0.588×10^{-3}**	0.515×10^{-3}	0.068×10^{-3}
Rank	2.1	1.8	2.1	2.3	2.0	1.8

Table 10
Top ranked negative and positive indicator terms (right).

Top ranked negative	Top ranked positive
mark-down	Neyt
Fitch	million
VFB	ticket
Humo	alfacam
rating	cable
restoration	stations

5.3. Explaining the models

The results in the previous sections suggest that choosing a bag-of-words model based on the full text of an article with a lag of 4 min is a reasonable choice, based on both the superior AUC and average Sharpe ratio. Given the fact that most of the results have substantial amounts of noise, we would like to validate whether the decisions that the model makes, are sensible or simply due to chance i.e. *how can we gain insight into text-mining-based stock-prediction models? (RQ3)*

5.3.1. Weights identification

The traditional approach to gaining insight into highly dimensional linear models is to look at the weight of each of the individual terms of the model (**w** in Eq. (3)). A higher weight of a term implies that the term has a higher influence on the resulting decision, should it occur. The weight of the top five ranked terms of the model are displayed in Table 10. The top-ranked negative results intuitively make sense (e.g., the impact of *Fitch rating*s during the period of our dataset was disastrous for many stock symbols and the Flemish Federation of Investorclubs and Investors (*VFB*) had a negative opinion on the macro-economical situation in Belgium). The positive ones, although arguably more far-fetched make sense as well: *Neyt* is one of the most important financial lawyers in Belgium, known for his expertise in pension funds and a *ticket* was usually used in the sense of gaining or losing a ticket to the stock exchange market. This effect is also a residual of the sparseness and high dimensionality of the data. As can be seen from Fig. 6, a marginal amount of terms receives a substantial weight in the model, whereas the bulk of terms receives a very small weight.

We can observe from the top words that they are contextual in time. Hence, one must ensure to monitor the model performance throughout time (this is known as backtesting in the finance literature Castermans, Martens, Van Gestel, Hamers, & Baesens (2010)). We can then retrain the model whenever necessary, including up-to-date information. Does this table of top-ranked words provide enough evidence for validating any individual decision made by the model? We believe it does not: many articles will not contain any of the words contained in the table (even if we were to expand it).

5.3.2. Explaining documents' classification

Martens and Provost (2014) argued that in document classification, the words in the individual explanations for classification decisions for specific documents vary tremendously. Their recently developed Explaining Documents' Classification

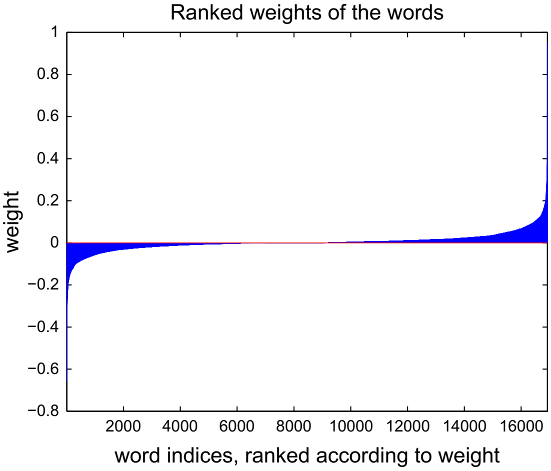

Fig. 6. The size of the weights for all of the 16.925 terms in the dataset ordered by weight.

(EDC) technique allows to look into why a model classifies an individual document as belonging to its predicted class in the form of an *explanation*, defined as a minimal set of words such that removing all words within this set from the document changes the predicted class.

Given a classified document (Extract 1, Appendix), EDC explains why that specific document was classified as an indicator for a positive or a negative trend. For this specific document, the explanation is shown in Explanation 1 in Appendix A. This article is classified as having a positive impact due to the fact that it contains words like *dividend* which the model learned to associate with positive target labels. Note that this word was only ranked at position 5453 (out of a total of 16925 words) and would never have shown up in the top-ranked weights table. The explanation makes sense to a human. For this document, we can therefore conclude that the model made a sensible (i.e. correct) decision. If it did not, we could use the information provided to us in a feedback loop to correct the model Fig. 2.

Note that EDC is not only useful for model validation, but could also be included as a feature in a trading dashboard where a human interprets the output of the trading model. This can be helpful in situations where words take on different meanings in different contexts, e.g. consider the word “window” which might be interpreted as being a part of a building or as the software name “Windows”. Differentiating between the two meanings is very difficult from an algorithmic point of view, but trivial to do for the operator of the trading tool.

5.4. Discussion

We have considered several measures to determine whether the model performs better than random or not. For the model considered in Section 5.3, all of the above tests give strong evidence for the fact that the model is indeed doing something more than random guessing, explaining its better than random performance. This is something that was not easily captured in a single aggregate measure such as accuracy or a statistical test (as is the current practice in the literature). We therefore argue that in future empirical work, trading models should be verified using more metrics and similar techniques whenever possible and applicable. During the operating phase of the models, similarly to standard credit scoring practices (Castermans et al., 2010), backtesting is advised in order to ensure the continuing correctness of the models.

Given the difficulties in assessing performance, we believe that providing lucid explanations of the inner workings of the models is quintessential to determining the reliability of decisions made by the models. Moreover, prior research has shown that the ability to explain a decision support model is of crucial importance for the acceptance of such models (Gregor & Benbasat, 1999; Kayande, De Bruyn, Lilien, Rangaswamy, & van Bruggen, 2008; Martens & Provost, 2014). We therefore recommend to use these highly technical models in combination with the interpretations of explanatory techniques. The implication of the previous statement is that these models should be used in a decision support tool, as opposed to a decision making tool. Unfortunately, the explanatory techniques previously discussed are of no use for these models since these only contain aggregate features. Thus, even though the global picture of the results displayed in Table 9 might look better at first sight, we would recommend using the individual models instead for reasons of transparency.

6. Conclusion

We have built several models that forecast stock price movement directionality based on news data, their sentiments and technical indicators. By using state-of-the-art explanation techniques, we have shown how to validate that these can perform slightly better than simple random guessing. We caution researchers to avoid the use of a single measure and strongly advise future research to use similar techniques to go beyond a simple aggregate performance metric to validate trading models, especially when attempting to provide counter-evidence for the Efficient Market Hypothesis.

During the operating phase of the models, we recommend using these highly technical models and easily interpretable explanation techniques in a decision support tool, as opposed to a decision making tool. Furthermore constant backtesting is advised in order to ensure to continuing correctness of the models.

In future research, we would like to expand our techniques to include even more variables and apply it to other markets with new dynamics as well. Particularly, we would like to investigate more detailed sentiment models based on the sentiment of verbs and nouns in addition to adjectives. In order to further develop the operational feasibility of trading models as decision support tools, we would additionally like to build such a tool (based on the presented research) and use it on a live stock market.

Appendix A. EDC Example

Extract 1: bag-of-words, full text, 4 minute lag.

Governement receives 400 million *euros* dividend from Belgacom

The *shareholders* of Belgacom on Wednesday at the annual *shareholders' meeting* approved a *dividend* of 2.18 *euros* per share for the year 2011. The Belgian government, majority *shareholder* of the telecommunications company, will therefore receive 394 million in *dividends*.

The dividend of 2.18 *euros* is identical to the *dividend* paid for 2010. In December 2011 Belgacom already paid an interim *dividend* of 0.50 *euros* per share in late April following the rest. The shareholders also approved the remuneration report.

In early March there was still great commotion when Belgacom announced no superdividends would be payed. The government had hoped it to be so since the government has 53.3 *percent* of all Belgacom shares in hands, and thus earns *significantly* whenever Belgacom pays a *dividend*. In an era full of budget constraints, an *extra dividend* to the Government would have been very good for them.

Departure-bonus for Concetta Fagard?

During the meeting, an individual shareholder went back to the 'Concetta Fagard'-affair concerning former assistant chief executive *Didier Bellens*. The shareholder wanted to know what Fagard received as departure bonus. The top of the group did not address that question, because it *belongs* to his 'private affairs'.

Explanation 1: bag-of-words, full text, 4 minute lag.

DECISION: buy and resell after 4 minute
REASON: explaining document with 69 features and class −1 (score −0.00257704) class changes:
Iteration 2 (from score −0.00257704 to 0.00135551)
→ IF (*dividend shareholders'-meeting*) are removed
Iteration 2 (from score −0.00257704 to 0.000209579)
→ IF (*dividend early*) are removed
Iteration 2 (from score −0.00257704 to 5.06409e-05)
→ IF (*dividend belong*) are removed
Iteration 2 (from score −0.00257704 to 0.00157765)
→ IF (*dividend bellens*) are removed
Iteration 2 (from score −0.00257704 to 0.000762401)
→ IF (*dividend didier*) are removed
Iteration 2 (from score −0.00257704 to 0.000401833)
→ IF (*euro dividend*) are removed
Iteration 2 (from score −0.00257704 to 9.19686e-05)
→ IF (*extra dividend*) are removed
Iteration 2 (from score −0.00257704 to 0.000454621)
→ IF (*significant dividend*) are removed
...

References

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B., & Belyaeva, J. (2010). Sentiment analysis in the news. In *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*.

Bodie, Z., Kane, A., & Marcus, A. (2008). *Investments*. McGraw-Hill.

Bollen, J., Mao, H., & Zeng, X.-j. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 1–8.

Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). Classification and regression trees.

Castermans, G., Martens, D., Van Gestel, T., Hamers, B., & Baesens, B. (2010). An overview and framework for PD backtesting and benchmarking. *Journal of the Operational Research Society, 61*, 359–373.

Cohen, A., & Hersh, W. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*.

De Smedt, T., & Daelemans, W. (2012b). Vreselijk mooi! (terribly beautiful): A subjectivity lexicon for Dutch adjectives. In *Proceedings of the 8th language resources and evaluation conference (LREC'12)* (pp. 3568-3572).

De Smedt, T., & Daelemans, W. (2012a). Pattern for python. *Journal of Machine Learning Research, 13*, 2063–2067.

Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. *Proceedings of LREC 2006*, 417–422.

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

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine, 17*(3), 37–54.

Gidofalvi, G. (2001) Using news articles to predict stock price movements.

Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly, 23*(4), 497.

Junqué de Fortuny, E., De Smedt, T., Martens, D., & Daelemans, W. (2012). Media coverage in times of political crisis: A text mining approach. *Expert Systems with Applications, 39*(14), 11616–11622.

Kayande, U., De Bruyn, A., Lilien, G. L., Rangaswamy, A., & van Bruggen, G. H. (2008). How incorporating feedback mechanisms in a DSS affects DSS evaluations. *Information Systems Research, 20*(4), 527–546.

Lavrenko, V., Schmill, M., Lawrie, D., & Ogilvie, P. (2000). Language models for financial news recommendation. *Proceedings of the Ninth International Conference of Information and Knowledge Management*.

Li, X., Wang, C., Dong, J., & Wang, F. (2011). Improving stock market prediction by integrating both market news and stock prices. *Lecture Notes in Computer Science: Database and Expert Systems Applications, 6861*, 279–293.

Malkiel, B. G. (1985). *A random walk down wall street*. W.W. Norton & Company.

Malkiel, B. G. (2005). Reflections on the efficient market hypothesis: 30 Years later. *The Financial Review, 40*(1), 1–9.

Martens, D., & Provost, F. (2014). Explaining data-driven document classifications. *MIS Quarterly, 38*(1), 73–99.

Mihalcea, R. (2007). The text mining handbook: Advanced approaches to analyzing unstructured data Ronen Feldman and James Sanger (Bar-Ilan University and ABS Ventures) Cambridge, England: Cambridge University Press, 2007, xii+410 pp; hardbound, ISBN 0-521-83657-3, 70.00. *Computational Linguistics, 34*(1), 125–127.

Mittermayer, M.-A. (2006). Newscats: A news categorization and trading system. *ICDM 2006*, 1002–1007.

Peramunetilleke, D. (2002). Currency exchange rate forecasting from news headlines. *Australian Computer Science Communications, 24*(2), 131–139.

Pui Cheong Fung, G., & Xu Yu, J. (2003). Stock prediction: Integrating text mining approach using real-time news. *IEEE International Conference on Computational Intelligence for Financial Engineering*, 395–402.

Rada, R. (2008). xpert systems and evolutionary computing for financial investing: A review. *Expert Systems with Applications, 34*(4), 2232–2240.

Schumaker, R. P., & Chen, H. (2009a). A quantitative stock prediction system based on financial news. *Information Processing & Management, 45*(5), 571–583.

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

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 56.

Suh, J. H., Park, C. H., & Jeon, S. H. (2010). Applying text and data mining techniques to forecasting the trend of petitions filed to e-People. *Expert Systems with Applications, 37*(10), 7255–7268.

Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications, 36*(7), 10760–10773.

Thomas, J. (2000). Integrating genetic algorithms and text learning for financial prediction. *Data Mining with Evolutionary Algorithms*.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. *Statistics for engineering and information science* (vol. 8). Springer.

Wilder, J. (1978). *New concepts in technical trading systems*. Trend Research, Greensboro, N.C.

Wuthrich, B., & Cho, V. (1998). Daily stock market forecast from textual web data. *IEEE International Conference on Systems, Man, and Cybernetics*, 1–6.

Zhai, Y., Hsu, A., & Halgamuge, S. K. (2007). Daily stock price trends prediction. In *Fourth international symposium on neural networks, dynamic system and control group* (pp. 1087–1096). University of Melbourne.