DEPARTMENT OF ENVIRONMENT,
TECHNOLOGY AND TECHNOLOGY MANAGEMENT

Design and Evaluation of Empirical Models for Stock Price Prediction

**Enric Junqué de Fortuny, Tom De Smedt,
David Martens & Walter Daelemans**

UNIVERSITY OF ANTWERP
Faculty of Applied Economics

Stadscampus
Prinsstraat 13, B.226
BE-2000 Antwerpen
Tel. +32 (0)3 265 40 32
Fax +32 (0)3 265 47 99
http://www.ua.ac.be/tew

FACULTY OF APPLIED ECONOMICS

DEPARTMENT OF ENVIRONMENT,
TECHNOLOGY AND TECHNOLOGY MANAGEMENT

Design and Evaluation of Empirical Models for Stock Price Prediction

Enric Junqué de Fortuny, Tom De Smedt, David Martens & Walter Daelemans

RESEARCH PAPER 2012-017
SEPTEMBER 2012

University of Antwerp, City Campus, Prinsstraat 13, B-2000 Antwerp, Belgium
Research Administration – room B.226
phone: (32) 3 265 40 32
fax: (32) 3 265 47 99
e-mail: joeri.nys@ua.ac.be

The papers can be also found at our website:
www.ua.ac.be/tew (research > working papers) &
www.repec.org/ (Research papers in economics - REPEC)

D/2012/1169/017

Design and Evaluation of Empirical Models for Stock Price Prediction

Enric Junqué de Fortuny*

Faculty of Applied Economics, University of Antwerp, Prinsstraat 13, B-2000 Antwerp, Belgium

Tom De Smedt

Faculty of Arts, University of Antwerp, Prinsstraat 13, B-2000 Antwerp, Belgium

David Martens

Faculty of Applied Economics, University of Antwerp, Prinsstraat 13, B-2000 Antwerp, Belgium

Walter Daelemans

Faculty of Arts, University of Antwerp, Prinsstraat 13, B-2000 Antwerp, Belgium

Abstract

The efficiënt market hypothesis and related theories claim that it is impossible to predict future stock prices. Even so, empirical research has countered this claim by achieving better than random prediction performance. Using a model built from a combination of text mining and time series prediction, we provide further evidence to counter the efficient market hypothesis. We discuss the difficulties in evaluating such models by investigating the drawbacks of the common choices of evaluation metrics used in these empirical studies. We continue by suggesting alternative techniques to validate stock prediction models, circumventing these shortcomings. Finally, a trading system is built for the Euronext Brussels stock exchange market. In our framework, we applied a novel sentiment mining technique in the design of the model and show the usefulness of state-of-the-art explanation-based techniques to validate the resulting models.

Keywords: stock prediction; Support Vector Machine; text mining; opinion mining

Highlights

- A stock prediction model based on opinion mining is built.
- Different metrics for evaluating prediction models are discussed.
- A thorough comparison with traditional modelling approaches is performed.
- Explanatory techniques are applied to gain insights into the model's decisions.

*Corresponding author

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 a great deal since those days. Nevertheless, stock price prediction remains an attractive topic for both researchers and investors[1]. Even so, during the past decades different theories have been developed to motivate why stock price prediction is not feasible per se. Recent studies have suggested empirical counter-arguments. This paper belongs to this category of empirical studies.

In order to do so, we build various empirical models and verify their performance on real life datasets by combining text mining techniques with time series prediction on articles from Belgian on-line news sites. Furthermore, we want to elaborate on how to evaluate such models. In the literature we encountered a large variety of evaluation metrics with many studies basing their conclusions solely on one or two metrics (e.g., accuracy, see Table 1). As we will see, one metric never contains all information and including additional evaluation methods might increase insight into the validity of the build models. We start by discussing popular theories and techniques related to stock prediction, followed by a more detailed look into text mining. Next, an empirical study and framework is elaborated on, with a deep analysis of the potential results.

1.1. Stock prediction theories

The *efficiënt market hypothesis* was first introduced in Fama (2012) and posits that the financial markets are informationally efficient. This implies that one can not 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. There are three main instances of the theory. The *weak* form states that only previous prices and historical information is incorporated in the current trade price of assets. The *semi-strong* form on the other hand, expands the information to that which is presently available as well. The last and strongest form claims that all information (even that which is not visible to the public) is included in the price at all times.

Random Walk Theory, developed by Malkiel (1985), states that the stock market prices of assets evolve in a pattern comparable to that of a Random Walk. The implication of this statement is that stock prices can not be predicted better than a chimpanzee throwing darts blindfold at a numerical scale board. Random Walk Theory is compatible with the efficient market hypothesis. Since no short trading system can do better than random predictions, the best trading strategy under this hypothesis is a long-term trading strategy of index funds.

Many economists, mathematicians and data mining practitioners believe that it is at least partly possible to predict stock market prices better than random predictions would. This is also visible in most frequently used analytical investment-tools, which use information based on financial data of the company (fundamental analysis) complemented with a pattern detection component that encodes evolutions in price- and volume-oscillations of stock commodities in a technical analysis (Malkiel, 2005).

The behaviour explained by both theories is based on the assumption that investors act rationally. In this publication we want to assess whether we can find evidence for the contrary: do investors sometimes show bias in their actions? The way in which we tackle this question is by extracting variables that could

[1] A quick google query for stock prediction reveals over 1.9 million results, including 1360 scientific publications

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

Table 1: Literature overview of news analysis for stock prediction.

be useful in indicating a bias from news published in popular media. The reason why this could work in a high frequency environment is that although stock-related news articles are ubiquitous, some human intervention is usually required in order to analyse them. This creates a lag between the appearance of an article and the trading action of the reader. Automatic trading systems could outperform human reaction in terms of speed, and thus in terms of revenues as well. However, as we will see, it is very important that proper evaluation is performed in order to determine the performance of a system.

1.2. Empirical Research on Stock Prediction

1.2.1. Empirical Stock Prediction

Many empirical models haven been built to perform directional predictions of stock movement based on text, with varying performances. These models assume that most if not all of the information can be found in the text itself. Li et al. (2011), however, remarked that this assumption is too narrow in that it disregards all other available 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 could stay upward (see Figure 1). In this case a negative directional prediction would be wrong, although the impact of the message itself was negative. An overview of some noteworthy previous empirical research is shown in Table 1, together with some extra features that we will discuss later.

The aforementioned behaviour can be remedied by adding proxies for the trend of the stock price and using those as well in the prediction model. Typically, one or more technical indicators are included. In this study we take a similar approach, but include two factors: a series of technical indicators and the sentiment of the news message (which hopefully induces a bias of the investors). It should be noted that it is very difficult to extract the exact effect of each variable individually ex post. Nevertheless it is not difficult to combine these in an empirical model as we will see in Section 3.

1.2.2. Evaluation metrics

As can be seen from Table 1, many different measures have been used to evaluate the performance of trading systems. Unfortunately, most of these do not give an accurate representation of the usefulness of the trading model by themselves.

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 to some

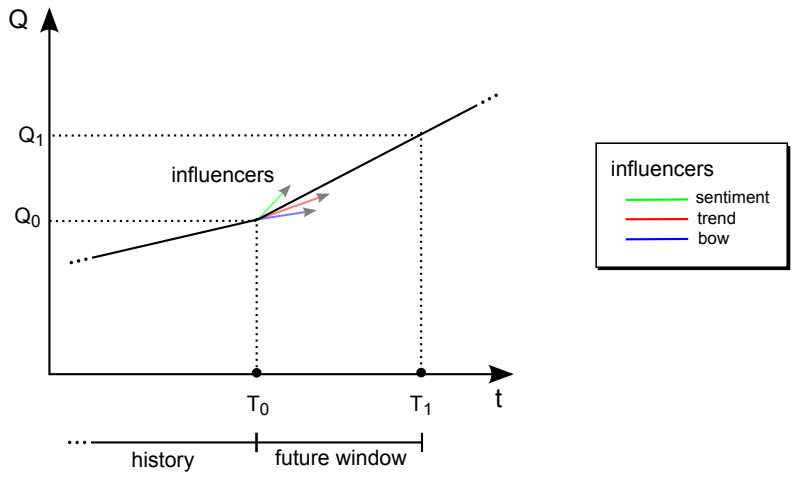

Figure 1: Many influences together determine the actual price as opposed to one variable.

extent. The problem with using accuracy as a performance measure is that it does not clarify well whether the model built actually performs better than random, due to the fact that the data in the test set is generally not evenly 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%, a good result at first sight, but not a good indication of an intelligent trading model. Note that many other similar measures suffer from this as well (including precision, lift and F scores)[2].

Discriminative power. In order to evaluate the discriminative power of the model, we proposed to use the Area Under the Receiver Operating Characteristic Curve metric (AUC, Fawcett (2006)), which gives the probability that the model will classify a randomly chosen positive instance higher than a randomly chosen negative one. The AUC is a generally accepted performance metric to assess the predictive performance of classification models in data mining. One of the advantages of using AUC is that it can cope with skewed distributions of target label data. Furthermore, it allows for easy comparison with random predictions (i.e., Random Walk Hypothesis), since a random classifier should result in a AUC value of exactly 50%.

We should note that all AUC and accuracy 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.

AUC is a useful metric for measuring model discrimination power. Even so, just like accuracy, it proves to be less useful in the real world evaluation of a classifier since it makes assumptions about the data that might not be portable to a real setting (particularly that the cost of a false negative and a false positive are equal). Alternative measures that simulate how the model is used in a real world setting do exist. We will discuss both a proxy for the profit of the trading simulation and the Sharpe ratio.

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

In our trading simulation, 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 some chosen trading strategy. In these approaches, the trading cost is usually assumed to be zero. The problem with these trading assessments is that it is very hard to compare these results to other publications since many factors can influence the outcome (e.g., starting budget, test set size, ...). A better way to evaluate profits is to use some form 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, one of which being that it does not take into account the actual risk undertaken by trading upon the built system. In Malkiel (2005), it is suggested that one of the main arguments to support the Random Walk Theory is 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 must therefore test our models in a trading simulation using a risk-weighted evaluation metric, discussed next.

Trading viability. In finance, risk is often defined in terms of variance of yields though one should 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). 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.

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

In order to properly evaluate the prediction performance of a model, no training information may 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 may not use any future information in the training phase of a model, this is usually referred to as *in-time*. A famous example that violates both of these principles received a lot of media attention recently. In a publication by Bollen et al. (2011), a stock prediction model is built, based on twitter mood prediction. Even though the study contains many methodological

and representational flaws[3], a $40 million hedge fund was started based on the technique, receiving a nomination for the 9th annual Awards for Excellence in Trading and Technology Europe 2011 for the most innovative trading Firm[4].

1.3. Text mining and sentiment analysis

1.3.1. Text mining

Text mining concerns the process of automatically extracting novel, non-trivial information from unstructured text documents (Fayyad et al., 1996), by combining techniques from data mining, machine learning, natural language processing (NLP), information retrieval (IR) and knowledge management (Mihalcea, 2011). 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 et al., 2010), guiding financial investments (Rada, 2008) and sentiment detection (Tang et al., 2009; Junqué de Fortuny et al., 2012).

In this particular context, we look for patterns in either occurring terms or the sentiment of the message that have an influence on the stock market price of a commodity. Typically, only the direction 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 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).

1.3.2. Sentiment analysis

Textual information can be broadly categorized into two types: objective facts and subjective opinions Liu (2010). Opinions carry people's sentiments, appraisals and feelings toward the world. Sentiment analysis (or opinion mining) is a subfield of natural language processing that in its more mature work focuses on two main approaches. The first approach is based on subjectivity lexicons, dictionaries of words associated with a positive or negative sentiment score (polarity, Taboada et al. (2011)). Such lexicons can be used to classify phrases, sentences or documents as subjective or objective, positive or negative. The second approach is by using machine learning text classification (see for example Pang et al. (2002)). Resources for sentiment analysis are interesting for marketing or sociological research. For example, they can be used to study customer product reviews (Pang et al., 2002), electronic word-of-mouth (Jansen et al., 2009), informal political discourse (Tumasjan et al., 2010) and public mood (Gilad and de Rijke, 2006). Using a subjectivity lexicon for Dutch adjectives we analyse the general tone associated with news articles on stock symbols to aid in the prediction of stock prices.

2. Material and Methodology

The main set-up of the training and testing of our stock prediction model is displayed in Figure 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

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

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

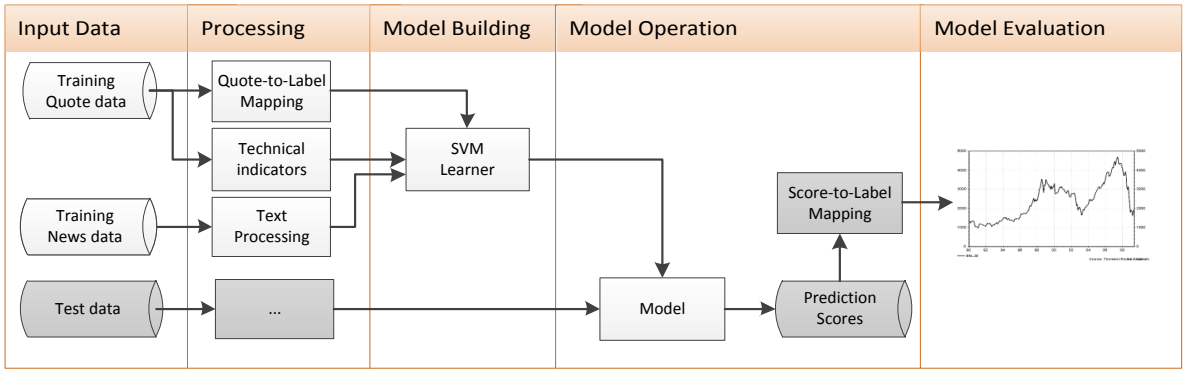

Figure 2: Main set-up of the model learning and evaluation procedure.

Source	#Readers[5]
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

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

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

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

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.

2.1. Input data

2.1.1. 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. 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. 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.

2.1.2. 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 $[w_1 w_2 \cdots w_m]$. 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:

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D) \tag{3}$$

$$\text{idf}(t, d, D) = \log \frac{|D|}{1 + |\{d \in D | t \in d\}|}, \tag{4}$$

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 usually don't contain discriminatory information. Empirical research has shown this to be a valuable transformation in numerous contexts Sebastiani (2002). Note that this scaling is of course based on information from the training set only.

2.1.3. Document sentiment polarity

For sentiment analysis, we used the previously created Pattern module for Python (De Smedt and 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

[5]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

module includes an algorithm that refines the score for each adjective by looking at preceding adverbs (e.g., extremely 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 and 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 in 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. The following assessment scores -0.17 for example:

"De grootste stroomproducent van Belgie dreigt ermee investeringen stil te leggen als de nucleaire taks wordt opgetrokken. 'Als de Belgische staat de door haar genomen engagementen niet zou nakomen, zou dit GDF Suez verplichten om haar beleid op het vlak van investeringen, tewerkstelling, opleiding en mecenaat in Belgie globaal te herzien.' GDF Suez-Electrabel, veruit de grootste stroomproducent in Belgie, lost een zwaar schot voor de boeg van de federale regeringsonderhandelaars."

"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."

2.1.4. Technical indicators

We considered four popular technical indicators in our approach, all of which are based on a series of price ticks $P = \{p_1, p_2, ..., p_n\}$ leading up to the last known price p_n (Bodie, Z. and Kane, A. and Marcus, 2008). The length of the series was chosen to be $n = 5$ ticks, which for the interday setting corresponds with a period of a week. For the intraday setting, we kept the amount of ticks, but sampled at higher frequency as well (30 minutes and 5 minutes).

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

$$RSI(P) = 100 - \frac{100}{1 + RS(P)} \tag{5}$$

$$RS(P) = \frac{\text{average gain}}{\text{average loss}} \tag{6}$$

According to Wilder (1978), the creator of this technical indicator, a stock pice 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.

$$R(P) = \frac{p_n - \min_{p_i \in P} p_i}{\max_{p_i \in P} p_i - \min_{p_i \in P} p_i}, \tag{7}$$

Psychological Line (PSY). 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:

$$PSY(P) = 100 \times \frac{|\{p_i | p_i \in P \wedge p_i > p_{i-1}\}|}{|P|}. \tag{8}$$

Bias. the bias indicator assesses the behaviour 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 behaviour 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:

$$BIAS(P) = 100 \times \frac{p_n - m(P)}{m(P)} \tag{9}$$

$$m(P) = \frac{1}{|P|} \sum_{p_i \in P} p_i \tag{10}$$

2.2. Target data

2.2.1. 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.

2.2.2. Quote-to-Label Mapping

In order to simplify the learning system, the problem is simplified to that of predicting a subset of possible stock movements, aggregated in classes (e.g., increase (+), stable ($\pm$), 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 as well. 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:

$$\Delta_r = \frac{Q_1 - Q_0}{Q_1} \tag{11}$$

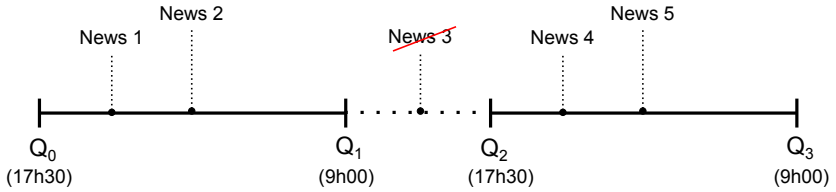

Figure 3: Time window and relative quote definition.

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 interday setting, the labels are based on opening and closing quotes. For the intraday setting, we tried several time windows after the appearance of an article (on a logarithmic scale, with a lag l ranging from 2^0 to 2^6 minutes), 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 (Figure 3). Previous studies have tried interpolating the values, but we believe this is a very rough approximation at best when working on a minute granularity level.

2.3. Support Vector Machines

The Support Vector Machine is a learning procedure based on the statistical learning theory Vapnik (1995). Given a training set 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. This hyperplane, $\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}) + b = 0$, 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

$$y(\mathbf{x}) = \text{sign}\left(\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}) + b\right), \tag{12}$$

however, it is never evaluated in this form. To solve the system of inequalities, it is reformulated as a convex optimization problem and solved using Lagrange multipliers. The exact details of this procedure are beyond the scope of this report but at the end this leads to the following classifier:

$$y(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i\, y_i\, K(\mathbf{x}_i, \mathbf{x}) + b\right), \tag{13}$$

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:

$$\begin{aligned}
&K(\mathbf{x}, \mathbf{x}_i) = \mathbf{x}_i^T\mathbf{x}, && \text{(linear kernel)}\\
&K(\mathbf{x}, \mathbf{x}_i) = \left(1 + \frac{\mathbf{x}_i^T\mathbf{x}}{c}\right)^d, && \text{(polynomial kernel)}\\
&K(\mathbf{x}, \mathbf{x}_i) = e^{\frac{-\|\mathbf{x}-\mathbf{x}_i\|_2^2}{\sigma^2}}, && \text{(RBF kernel)}\\
&K(\mathbf{x}, \mathbf{x}_i) = \tanh(\kappa\,\mathbf{x}_i^T\mathbf{x} + \theta), && \text{(MLP kernel)},
\end{aligned} \tag{14}$$

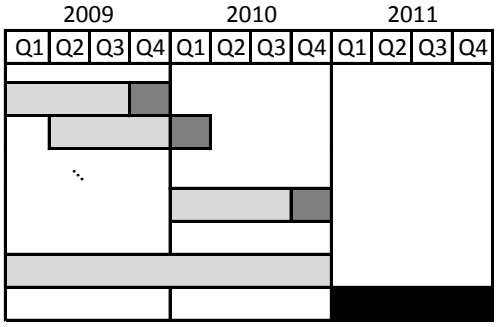

Figure 4: Train/test data split for the interday setting.

where d, c, σ, κ and θ are constants. Note that for the MLP kernel, the Mercer condition is not always satisfied. Throughout this study we will be using a linear kernel due to the high dimensionality of the data and speed considerations.

2.4. Evaluation

In order to ensure good generalization properties, the SVM uses separate datasets for training, validation and testing (as displayed in Figure 4). The testset remains untouched in 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 white), and then evaluating it on another part (the validation set, displayed in gray). By repeating this process over all five folds, we 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 in any step of the process 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 1.2.2, we will include four evaluation metrics in this study: accuracy, AUC, return and Sharpe ratio.

3. Results

3.1. Individual model training

In a first batch of experiments we built 56 models for each of the types of features on the eight different minute lags. For both the bag-of-words and the sentiment we considered three different partitionings of the texts (as previously explained in Section 2.1.1).

3.1.1. Prediction accuracy

The accuracies of each of the models is displayed in Table 4. Some models seem to perform well, others don't. The results are quite similar to those of other empirical studies, which go on to conclude that some of the models are better than random. As discussed before it's very hard to tell from these numbers which model performed better than random since the distribution of positive and negative

	BOW			Sentiment			TI
Lag t	Full text	Paragraph	Title	Full text	Paragraph	Title	
1m	50.00%	46.13%	54.78%	57.37%	43.77%	69.05%	47.19%
2m	58.43%	44.35%	55.06%	61.22%	60.38%	29.76%	59.83%
4m	53.65%	51.19%	43.26%	57.69%	56.55%	34.52%	57.58%
8m	53.09%	53.27%	46.63%	54.81%	46.01%	44.05%	55.34%
16m	52.81%	53.87%	49.72%	50.96%	49.84%	40.48%	55.06%
32m	46.63%	50.60%	51.97%	50.32%	49.20%	63.10%	46.07%
64m	45.51%	53.57%	50.28%	45.83%	46.01%	50.00%	46.91%
24h	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

Table 4: Prediction ability (Acc) of the trained models on the test set.

	BOW			Sentiment			TI
Lag t	Full text	Paragraph	Title	Full text	Paragraph	Title	
1m	48.70%	48.69%	55.18%	50.07%	49.77%	53.81%	52.87%
2m	49.03%	47.33%	51.19%	48.99%	49.26%	43.08%	56.03%
4m	53.30%	50.19%	40.93%	47.62%	47.67%	39.25%	55.35%
8m	56.82%	55.16%	47.12%	48.63%	51.72%	43.44%	58.22%
16m	54.28%	54.78%	46.75%	48.66%	52.11%	49.35%	54.91%
32m	51.36%	49.11%	50.99%	46.36%	45.66%	57.33%	51.27%
64m	45.90%	52.60%	49.25%	44.98%	43.64%	39.12%	53.20%
12h	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

Table 5: Discrimination ability (AUC) of the trained models on the test set.

samples in the test set is skewed. One way to solve this issue would be to make sure that it is uniformly distributed with an equal 50% chance for each of both classes. We would like to point out that even then, any of the above 50% values could have been attained merely by chance. That is, as the number of experiments performed increases, so does the chance of attaining seemingly better than random result.

3.1.2. Area Under Curve

The AUC values of all models over the testset are displayed in Table 5 for the bag-of-words (BOW), sentiment and technical indicator (TI) approach. Using AUC it is much easier to compare our models to a random prediction model since a random prediction model has an AUC of exactly 50%. From Table 5 we can discern that, according to AUC, it is certainly not always possible to build a model that performs better than random predictions. Nevertheless some interesting results do stand out. The technical indicators perform very well, confirming that they could indeed be useful for predicting stock price behaviour. This comes as no surprise since technical indicators have been used in trading tools. Technical indicators perform significantly better than all other models and random selection except for the sentiment-title model (using a Wilcoxon signed rank test, significance level $p < 0.05$).

The sentiment results are inconsistent and often underperform the other models as compared to the BOW approach or the technical indicators (often even underperforming to a random classifier). This is partly due to the fact that not much sentiment information was available (i.e., sometimes no sentiment was detected at all) and partly due to the fact that the sentiment extraction model was built on a dataset

	BOW			Sentiment			TI
Lag t	Full text	Paragraph	Title	Full text	Paragraph	Title	
1m	$0.185\,10^{-3}$	$0.205\,10^{-3}$	$0.189\,10^{-3}$	$0.000\,10^{-3}$	$0.195\,10^{-3}$	N.A.	$0.190\,10^{-3}$
2m	$0.273\,10^{-3}$	$0.287\,10^{-3}$	$0.283\,10^{-3}$	N.A.	N.A.	$0.496\,10^{-3}$	$0.284\,10^{-3}$
4m	$0.268\,10^{-3}$	$0.283\,10^{-3}$	$0.266\,10^{-3}$	N.A.	N.A.	$0.692\,10^{-3}$	$0.269\,10^{-3}$
8m	$0.098\,10^{-3}$	$0.105\,10^{-3}$	$0.119\,10^{-3}$	N.A.	$0.151\,10^{-3}$	N.A.	$0.105\,10^{-3}$
16m	$-0.053\,10^{-3}$	$-0.034\,10^{-3}$	$-0.066\,10^{-3}$	N.A.	N.A.	$0.432\,10^{-3}$	$-0.053\,10^{-3}$
32m	$-0.056\,10^{-3}$	$-0.012\,10^{-3}$	$-0.047\,10^{-3}$	N.A.	N.A.	N.A.	$-0.060\,10^{-3}$
64m	$-0.158\,10^{-3}$	$-0.108\,10^{-3}$	$-0.122\,10^{-3}$	N.A.	N.A.	N.A.	$-0.125\,10^{-3}$
24m	$0.277\,10^{-3}$	$0.321\,10^{-3}$	$0.278\,10^{-3}$	$0.156\,10^{-3}$	$0.158\,10^{-3}$	$0.231\,10^{-3}$	$0.165\,10^{-3}$
rank	5.1	2.6	4.3	4.4	3.8	3.4	4.5

Table 6: Average monthly return rates on the test set in percentage (buy all positives / maximum buy selected).

in a previous study in a different context. This result could therefore indicate that sentiment from one context is not portable to a different one. It may be possible that in contrast to the domains on which the sentiment extraction model was trained, 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, however, 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 this failure. 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.

3.1.3. Rate of Return

Table 6 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 behaviour is an indication of possible bad convergence during the learning procedure since it has a strong bias to classify news as being negative.

Interestingly though, all short trading time span systems ($\leqslant$ 8m) performed positively in terms of return. This corresponds to our original hypothesis: the behaviour that we attempt to capture using our models occurs in this time period since the original premises is that we want to predict how traders react to a news message. 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. Note, however, that one must watch out for data dredging 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 1.2.2), an important drawback of using returns is that it is very difficult to compare the resulting values, since not all of the above models have the same test-set size, furthermore, these values do not take into account the risk undertaken by the trader when performing these actions.

3.1.4. Sharpe ratio

The resulting Sharpe ratios for our simple trading simulation are displayed in Table 5, 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. Again, shorter trading time span systems (2m, 4m)

	BOW			Sentiment			TI
Lag t	Full text	Paragraph	Title	Full text	Paragraph	Title	
1m	$-1.484\,10^{-3}$	$-0.852\,10^{-3}$	$-1.354\,10^{-3}$	N.A.	$-1.140\,10^{-3}$	N.A.	$-1.344\,10^{-3}$
2m	$0.954\,10^{-3}$	$1.242\,10^{-3}$	$1.182\,10^{-3}$	N.A.	N.A.	$4.646\,10^{-3}$	$1.193\,10^{-3}$
4m	$0.711\,10^{-3}$	$1.003\,10^{-3}$	$0.676\,10^{-3}$	N.A.	N.A.	$8.008\,10^{-3}$	$0.730\,10^{-3}$
8m	$-2.202\,10^{-3}$	$-2.022\,10^{-3}$	$-1.863\,10^{-3}$	N.A.	$-1.351\,10^{-3}$	N.A.	$-2.087\,10^{-3}$
16m	$-3.658\,10^{-3}$	$-3.323\,10^{-3}$	$-3.817\,10^{-3}$	N.A.	N.A.	$2.645\,10^{-3}$	$-3.657\,10^{-3}$
32m	$-2.305\,10^{-3}$	$-1.899\,10^{-3}$	$-2.229\,10^{-3}$	N.A.	N.A.	N.A.	$-2.342\,10^{-3}$
64m	$-2.464\,10^{-3}$	$-2.081\,10^{-3}$	$-2.220\,10^{-3}$	N.A.	N.A.	N.A.	$-2.240\,10^{-3}$
12h	$0.293\,10^{-3}$	$0.344\,10^{-3}$	$0.293\,10^{-3}$	$0.157\,10^{-3}$	$0.159\,10^{-3}$	$0.257\,10^{-3}$	$0.168\,10^{-3}$
rank	5.6	3.1	4.8	3.1	4.3	2.1	5.0

Table 7: Average Sharpe ratio of the AUC-trained models on the test set (buy all positives / maximum buy selected).

perform better. 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 1m high frequency setting): although return rates were very high, the Sharpe ratios look bad. This indicates that the trading strategy for the interday settings are not without risk, something we could not have known by only looking at the return rates.

3.1.5. Explaining the models

The results in the previous section indicate that choosing a bag-of-words model based on the full text of an article with a lag of 4 minutes seems like a reasonable choice, based on both the superior AUC and average Sharpe ratio. Given the fact that most of the results contain a lot of noise, we would like to verify whether the decisions that the model makes, are sensible or simply due to chance.

Weights identification. The traditional approach to gain insight into high dimensional linear models is to look at the weight of each of the individual terms of the model. 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 Figure 5. 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 somewhat 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 Figure 5, a marginal amount of terms receives a substantial weight in the model, whereas the bulk of terms receives a very small weight.

As mentioned in the previous paragraph, we can observe from the top words that they are contextual in time, to avoid mistakes in the far future, one must therefore ensure to perform backtesting and expand the train set and retrain the model whenever necessary. Returning to the main reasoning: does this table 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). The next section discusses a way to circumvent this problem.

Explaining Documents' Classification. Martens and Provost (2011) 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 (EDC) technique allows to look

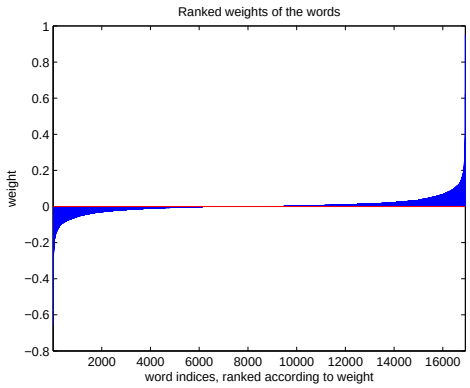

Figure 5: The size of the weights for all of the 16.925 terms in the dataset ordered by weight.

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

Table 8: Top ranked negative and positive indicator terms (right).

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 the Appendix. 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 is lucid and makes sense to a human interpreter. For this document, we can therefore conclude that the model made a correct and comprehensible decision. 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.

3.2. A hybrid model

In Section 1.2.1 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 reacted 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.

In order to give equal importance to each of the variables, the (very high dimensional) BOW is not maintained in it's full form, but only the scores of the BOW 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 for

Lag t	Accuracy Full text	Accuracy Paragraph	Accuracy Title	AUC Full text	AUC Paragraph	AUC Title
1m	55.38%	55.23%	65.38%	51.53%	48.27%	44.61%
2m	60.22%	60.47%	59.62%	52.69%	49.32%	44.44%
4m	61.29%	55.23%	59.62%	56.35%	39.23%	39.17%
8m	51.61%	43.60%	46.15%	48.20%	55.37%	57.93%
16m	46.77%	50.58%	57.69%	52.23%	51.02%	56.97%
32m	45.16%	44.19%	46.15%	54.62%	52.58%	53.07%
64m	47.85%	59.30%	53.85%	48.78%	43.41%	60.71%
12h	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		
1m	$0.083\,10^{-3}$	$0.131\,10^{-3}$	$0.099\,10^{-3}$	$-4.735\,10^{-3}$	$-3.353\,10^{-3}$	$-4.449\,10^{-3}$
2m	N.A	N.A	$0.160\,10^{-3}$	N.A	N.A	$-2.201\,10^{-3}$
4m	$0.267\,10^{-3}$	N.A	$0.075\,10^{-3}$	$0.716\,10^{-3}$	N.A	$-3.702\,10^{-3}$
8m	$0.141\,10^{-3}$	$0.212\,10^{-3}$	$0.127\,10^{-3}$	$-1.693\,10^{-3}$	$-0.401\,10^{-3}$	$-2.079\,10^{-3}$
16m	$-0.123\,10^{-3}$	$-0.021\,10^{-3}$	$0.025\,10^{-3}$	$-5.329\,10^{-3}$	$-3.792\,10^{-3}$	$-3.401\,10^{-3}$
32m	$-0.093\,10^{-3}$	$-0.035\,10^{-3}$	$-0.025\,10^{-3}$	$-3.519\,10^{-3}$	$-2.959\,10^{-3}$	$-2.517\,10^{-3}$
64m	$-0.675\,10^{-3}$	$-0.578\,10^{-3}$	$-0.254\,10^{-3}$	$-6.267\,10^{-3}$	$-5.746\,10^{-3}$	$-3.554\,10^{-3}$
12h	$0.282\,10^{-3}$	$0.252\,10^{-3}$	$0.056\,10^{-3}$	$0.588\,10^{-3}$	$0.515\,10^{-3}$	$0.068\,10^{-3}$
rank	2.1	1.8	2.1	2.3	2.0	1.8

Table 9: Hybrid AUC scores.

the intraday setting and 10 variables for the interday setting. 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 behaviour the faster acting models outperform the slower models with the exception of the interday model.

In terms of return rates and Sharpe ratio the models reach some high levels, but this is somewhat countered by the fact that some of the built models compare equal or even a little bit worse than the individual models. 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 for reasons of transparency recommend using the individual models instead.

3.3. Discussion

We considered several measures to determine whether the model performs better than random or not. For the model considered in Section 3.1, all of the above tests gave strong evidence for the fact that the model is indeed doing something more than simple 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, backtesting is advised in order to ensure to continuing correctness of the models.

Furthermore, we believe that no trading model can truly be determined to be reliable without some insights into the model as provided by an explanatory technique. We therefore recommend to use these

highly technical models and easily interpretable explanation 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.

4. 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 avoid the use of a single measure but to go beyond one performance metric 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 providing counter-evidence for the Efficiënt 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. Additionally, we would like to research other evaluation metrics and their relevance and impact on claims of better than random performance.

Appendix A. EDC Example

Extract 1: BOW, 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: BOW, full text, 4 minute lag

DECISION: buy and resell after 4 minutes
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., and Belyaeva, J. (2010). Sentiment Analysis in the News. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

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

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

De Smedt, T. and Daelemans, W. (2012a). Pattern for Python. *Journal of Machine Learning Research*, 13(2063-2067).

De Smedt, T. and 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)*, pages 3568–3572.

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

Fama, E. (2012). 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., and 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.

Gilad, M. and de Rijke, M. (2006). Capturing Global Mood Levels using Blog Posts. In *Proceedings of The Spring Symposia on Computational Approaches to Analyzing Weblogs (AAAI-CAAW)*.

Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60:2169–2188.

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

Lavrenko, V., Schmill, M., Lawrie, D., and 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., and 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.

Liu, B. (2010). Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing*, (1):1–38.

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. and Provost, F. (2011). Explaining Documents' Classifications. Working paper CeDER. Stern School of Business, New York University.

Mihalcea, R. (2011). 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*, pages 1002–1007.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. In *In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing EMNLP02*, pages 79–86.

Peramunetilleke, D. (2002). Currency exchange rate forecasting from news headlines. *Australian Computer Science*, 5.

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

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

Schumaker, R. P. and Chen, H. (2009). A Quantitative Stock Prediction System based on Financial News. *Information Processing & Management*, 45(5):571–583.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*.

Suh, J. H., Park, C. H., and 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.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 1(September 2010):1–41.

Tang, H., Tan, S., and 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*.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*, volume 8 of *Statistics for Engineering and Information Science*. Springer.

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

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

Zhai, Y., Hsu, A., and Halgamuge, S. K. (2007). Daily Stock Price Trends Prediction. pages 1087–1096.