An Automated Framework for Incorporating News into Stock Trading Strategies

Wijnand Nuij, Viorel Milea, Frederik Hogenboom, Flavius Frasincar, and Uzay Kaymak

Abstract—In this paper we present a framework for automatic exploitation of news in stock trading strategies. Events are extracted from news messages presented in free text without annotations. We test the introduced framework by deriving trading strategies based on technical indicators and impacts of the extracted events. The strategies take the form of rules that combine technical trading indicators with a news variable, and are revealed through the use of genetic programming. We find that the news variable is often included in the optimal trading rules, indicating the added value of news for predictive purposes and validating our proposed framework for automatically incorporating news in stock trading strategies.

Index Terms—Computer applications, evolutionary computing and genetic algorithms, learning, natural language processing, web text analysis.

1 Introduction

Financial markets are driven by information. An important source of information is news communicated by different media agencies through a variety of channels. With the increasing number of information sources, resulting in high volumes of news, manual processing of the knowledge being conveyed becomes a highly difficult task. Additionally, given that this information is time-sensitive, especially in the context of financial markets, selecting and processing all the relevant information in a decision-making process, such as the decision whether to buy, hold, or sell an asset is an especially challenging task. This environment motivates a need for automation in the processing of information, to the extent that investment decisions where the news factor plays an important role can be based on an automatically generated recommendation that takes into account all news messages relevant to a certain financial asset.

In previous work we have devised lexico-semantic patterns for information extraction from news that extend the well-known lexico-syntactic patterns with semantic aspects [1], [2]. Using information extracted from text in a financial context recently enjoys increasing attention. In [3] the authors extract investor sentiment from stock message boards. The prediction of bankruptcy of firms, as well as fraud, based on textual data from the Management Discussion and Analysis Sections (MD&A) of 10-K reports is investigated in [4]. A popular Wall Street Journal column is used for investigating asset prices and trading volumes in [5]. Financial news stories are used for the prediction of stock returns and firms' future cash flows in [6]. Thus, the qualitative data may emerge from different sources, and can be used for the prediction of different financial aspects of firms' performance.

We focus on information presented in textual format, i.e., financial news messages with a particular focus on companies listed under the FTSE350 stock index. The research question addressed is how the information communicated through textual news messages can be automatically incorporated into trading strategies. We use a three-step approach consisting of: (i) extracting the relevant events, as well as the involved entities, from the text of the news messages, (ii) associating an impact with each of the extracted events, and (iii) making use of the impact of news events in trading strategies.

Upon extracting the events and associating these with a predefined impact, trading rules based on news can be derived. We only consider technical trading indicators as part of these trading rules, but the approach can be easily extended to incorporate other indicators, e.g., those initiating from fundamental analysis. Technical trading has been used previously for financial forecasting [7], [8], thus motivating our choice for this approach. The constructed trading strategies are expressed as trees, where the leaves are technical indicators or news event indicators and internal nodes represent the logical operators 'and' and 'or'. These trading strategies generate a buy or sell signal for the assets they are applied to, and are determined through genetic programming where a pool of possible trading strategies is tested on historical stock data.

We hypothesize that, if the proposed framework is valid, news will be included in the trading strategies generated through genetic programming. Addition-

- *Wijnand Nuij is with Semlab, Zuidpoolsingel 14a, NL-2408 ZE Alphen a/d Rijn, the Netherlands. E-mail: nuij@semlab.nl*
- *Viorel Milea, Frederik Hogenboom, and Flavius Frasincar are with the Erasmus School of Economics, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, the Netherlands. Email: {milea,fhogenboom,frasincar}@ese.eur.nl*
- *Uzay Kaymak is with the Department of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology, P.O. Box 513, NL-5600 MB Eindhoven, the Netherlands. E-mail: u.kaymak@ieee.org*

ally, the trading strategies that we derive in this way should generate positive returns. The first hypothesis comes from the idea that, when providing a genetic program with a pool of variables without the restriction that all these variables should be included in a trading strategy, only the variables that are maximizing the returns will be selected. Trading strategies including a news variable will thus indicate that the content of the news messages has been quantified in a way that enables generation of profit beyond the ability of trading rules based solely on technical analysis. The second hypothesis states that, next to generating trading strategies based on news, the resulting rules should also be able to obtain a positive return.

The paper is structured as follows. First, we present previous work on the relationship between news and the stock market, and the type of events that are proven to influence stock prices. Next, we provide an initial, quantitative investigation of the relationship between news messages and the stock market. Subsequently, we discuss the technical indicators that we use for deriving stock trading strategies. After that we introduce our framework for automated trading based on news and discuss the results of validating the framework. Last, we give some practical considerations and conclude this paper.

2 Related Work

Regarding the relationship between news and the stock market, we consider three key aspects: (i) there is evidence that a relationship exists between news announcements and financial markets, (ii) the impact of events on financial markets can be quantified, and a list of relevant events can be identified, and (iii) the relationship between information in the form of news and financial markets is not a trivial one. One aspect left aside relates to mining news messages for assessing market response. For a survey of the different methods employed for this purpose, we refer the reader to [9].

In [10] the relation between the number of news announcements and trading activity is investigated. The research is focused on whether the amount of information that is publicly reported affects the trading activity and the price movements in the stock market. Here, information is defined and quantified by the number of daily announcements released by the Dow Jones & Company newswire. All news messages are assigned equal importance, regardless of the type of event being described. The results indicate a statistically significant positive correlation between the number of daily news announcements and trading activity. If the number of announcements increases with 100%, the trading activity will increase with 38%. The relation becomes stronger only if those news messages are selected that, besides being published through the wire service, are published in a newspaper the next morning.

The relation between news announcements and monthly returns is also investigated in [11]. Several stocks are selected with at least one news story in a certain month. The news messages are divided into 'news winners' (price increased after announcement) and 'news losers' (price decreased after announcement). The abnormal returns are measured for 36 months after the month when the news was published. The results are compared to a group 'no news' containing those companies which had no news in a certain month. The authors conclude that stocks exhibit abnormal returns after public news.

The effect of analyst recommendations, with a focus on buy advices, is studied in [12]. First, the authors test whether the advice issued especially for clients (before the opening of the stock market) contains information and then perform the same test but following the official release of the advice (to the main public). The authors find a strong relation between an initial coverage with a buy recommendation and a reaction in the stock market.

The relation between earnings announcements and trading volume around the announcement date is investigated in [13], focusing on AEX exchange stocks from 1994 to 1999. A significant positive increase in trading volume is found around an announcement. The increase in trading activity is the largest at the announcement date. The robustness of the relation is checked for small and large companies. Both categories have a significant relation with trading activity, but the relation is much stronger regarding small companies compared to large companies. A possible explanation is that there is less information available about small companies. Another relation was found between the date of the announcement and trading volume. The longer a company waits with revealing the earnings, the smaller the change in trading volume. A possible explanation is that the expectations are more accurate in that case, i.e., analysts have more time and information (earnings from competitors) to accurately predict earnings.

Up until now, different meanings have been assigned to the word news when studying the relation to the stock market. The employed news sources are arbitrary, they contain different news messages, and are not complete. Although a relation is apparent, it is necessary to zoom into real-life events and quantify the relation between these events and returns. The list of events that possibly affect the stock price is extremely large, but in the remaining paragraphs we focus on a limited number hereof, considered to be of increased relevance in financial markets.

A management change event is a change in the set of individuals holding the title Chief Executive Officer (CEO), president, or chairman of the board [14]. The reaction after a management change indicates whether the market considers this event as important. A stock return after a management change contains:

- The information effect (negative): the management performance is worse than expected by the market.
- The real effect (positive): the change is in shareholders' interest. If a company performs very bad, a management change could mean a new vision, strategy, etc., so the expectations about the companies' future results could be revised. The news is received positively.

The authors did not find a general relation between a management change and abnormal stock returns. Only on the day of the announcement a statistically significant price movement was noticed, but the direction could be both positive and negative.

The effect of a management change is studied in [15]. The identified average excess return from the day before until the day of an announcement is 2.479% (positive significant). Also, the title power, the company size, and the manager type have a positive, significant impact. Generally, a management change conveys bad news about the company's performance, but a management change is received positively if the company performance is bad.

The trading activity and price movements before, on, and after the day of a merger announcement are studied in [16]. In 79% of the acquired firms a significant increase in trading volume is found one week before the announcement, compared to 3 months before that date. Approximately half of the reactions occur before the official public announcement – they start one month before the merger. The strongest reaction in 1 day is on the announcement day itself: the market reacts immediately.

The price momentum following a merger announcement is investigated in [17]. Price momentum relates to an initial market response to a merger announcement and its propagation through time, i.e., if the initial reaction is positive, it tends to continue to be so. The results indicate that if a company is associated with successful mergers in the past, this will positively influence price momentum.

Stock splits and their effect on the price are studied in [18]. An NYSE sample from 1988 until 1997 is used with over 3000 stock splits. The authors find a 9% positive difference of abnormal returns between the split stocks and a control group, a year after the split.

The price reaction after dividend initiations and omissions is investigated in [19] for short term (3 days) and long term (several years) using a buy and hold strategy to measure returns. In the 3 days around an initiation announcement a significant excess return of 3.4% is found. In the year before, the excess return is 15.1%. Companies with a dividend omission perform very poor in the year before the announcement, apparent from an excess return of -31.8%. Around the announcement, an excess return of -3.1% is identified. These trends continue also in the next 1 year and the next 3 years after the announcement.

These findings come to support our assumption that events that can be identified in news messages have a significant impact on stock prices and trading volumes. For this reason, we consider it worthwhile to employ such events in our analysis. In the final part of this section we focus on different properties of (public) information in the context of financial markets.

The degree of uncertainty of information is explored in [20]. The hypothesis is that greater information uncertainty will lead to higher expected stock returns after good news on the one hand, and lower expected returns after bad news on the other hand. This implication is based on results from behavioural finance studies, i.e., psychological biases such as overconfidence are increased when there is more uncertainty. Here, good and bad news are defined as upward and downward analyst forecast revisions, respectively. Evidence is found that the market reaction directly after an announcement is incomplete, i.e., bad news implies relatively lower future returns and good news predicts higher future returns.

The influence of certain forms of rumours on trading activity on the stock market is evaluated in [21] through a dynamic model with three kinds of rumours: honest, bluffing, and cheating. It is concluded that spreading rumours makes economic sense. Rumours can increase stock demand and drive the price above the real price. In case of cheating with false rumours, followers will not use the trader's rumours, causing the rumourmonger to lose his reputation.

Our approach does not focus on representing and reasoning with complex knowledge contained in news messages, but rather focuses on single events. It would be possible to design an ontology of events in the Web Ontology Language (OWL) [22] if the context is static, although a temporal web ontology language such as tOWL [23], [24], [25], [26] should be more suited for representing and reasoning with the facts exposed by the news messages. However, as we focus on single events extracted from news, we do not rely on an approach based on ontologies.

Similar work regarding the extraction of optimal trading rules based on technical indicators related to price is presented in [27]. However, unlike the current research, news are not used in trading strategies. Also, in previous work [28] we have successfully used news events for financial risk analysis by improving the historical Value-at-Risk method.

Our current approach is novel in that it does not focus on a particular event type, but rather on a thesaurus of events that play a significant role in financial markets. Rather than focusing on news volume, we extract relevant events from market announcements and try to include them in trading rules. For this, we employ genetic programming that can choose between different variables in creating profitable trading rules. The variables originate in technical analysis, except for the news-related variable.

3 A Preliminary Analysis of the Relationship between News and the Stock Market

Our analysis of the relationship between news and the stock market, as apparent from the collected dataset is focussed on discovering the influence that news have on the share price of the concerned companies, as well as on whether this influence can be captured through the extraction of events from news messages and employing a predefined impact for determining the direction of this influence on prices.

3.1 Event Information Extraction

The event information extraction from the news messages is based on recognizing a predefined set of events as well as the affiliated entities. For this we rely on the ViewerPro tool (available at http://viewerpro.semlab.nl/pages/p.view?id=212), a proprietary application able to extract events from text-based data.

ViewerPro is an application created by SemLab that enables the identification of events in news messages. These events can be used to determine the impact of a news item on an equity. ViewerPro turns enormous amounts of unstructured news into structured trading information. Once the unstructured news information is fed in the ViewerPro system, it undergoes several (proprietary) processing steps in order to filter out unwanted information and select solely that which is relevant. Applied procedures are (amongst others) metadata filtering, parsing, gazetteering, stemming, and automatic pattern matching.

The ViewerPro system relies on a domain specific knowledge repository, i.e., an ontology with properties and lexical representations of financial entities (companies). First, concepts from the domain ontology are matched in incoming news items. Subsequently, using a proprietary heuristic based on semantical, morphological, syntactical, and typographical inputs, the list of concepts is segmented into groups of related concepts. Last, ViewerPro identifies events (predefined semantic concepts describing important message content) by means of pattern matching.

Large amounts of news messages are filtered for equity-specific news and the semantic analysis system of ViewerPro interprets the impact of every individual news message.

3.2 Descriptive Statistics of the Dataset

The dataset we employ consists of a database of historical company share prices as well as a collection of news messages related to these companies. The company dataset consists of all firms included in the FTSE350 stock index at August 1st, 2008. Stock prices are scraped from Yahoo! Finance ticker data. The news dataset is collected through the Reuters news feed, and concerns all 350 companies listed under FTSE350. Both datasets cover the period January 1st, 2007 until April 30th, 2007. The news dataset provides a set of 5,157 events. However, only a subset hereof is employed for our study. The selection of these relevant events is based on three criteria:

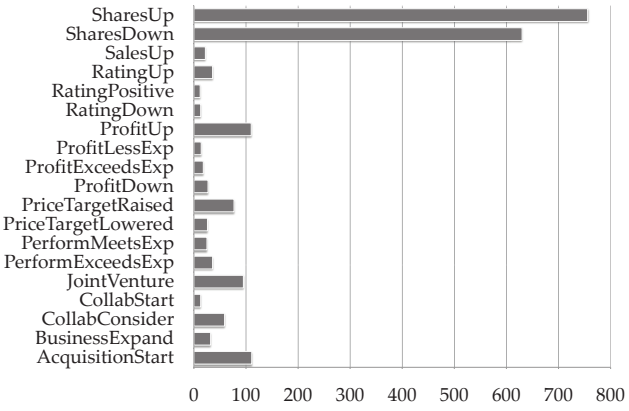

Fig. 1. Frequency of events in the dataset.

- News articles issued on days when the stock exchange is closed are not considered;
- Duplicate events are removed;
- Rare events ($< 0.5\%$ of all events) are omitted.

We do not include articles issued on days when the stock exchange is closed as the events contained in these messages will not have an immediately quantifiable impact on the stock price. Since several events can occur during the period when the stock exchange is closed, associating these events with changes in price over this period will introduce additional variance with regard to which event precisely influences the change in price.

At times, news messages may be repeated to provide updates on an event described in a previous message. This results in events that are on the same day, concern the same company, and are identical to another event previously extracted on the same day. Since these news messages describe the same events, it suffices to only consider them once and thus incorporate the associated impact for the event in the stock price projection only once.

Infrequently occurring events, i.e., events occurring in less than 0.5% of the news messages, are removed from the dataset as considering them would negatively influence the statistical validity of our conclusions. Moreover, the impact of such isolated events is difficult to assess with confidence.

Upon considering these four criteria on the event dataset we have collected, the original sample of 5,157 events is reduced to 2,112. An overview of these events, as well as their frequencies in the event dataset, is presented in Fig. 1.

3.3 Relationship between News and Share Prices

The impact of news on stock prices is assessed using relative returns, based on end-of-day data, i.e., closing prices P. For a single asset, a return is computed as:

$$r_i = \frac{P_{i+n} - P_i}{P_i} \times 100 \ , \tag{1}$$

where i represents the day before the event and n represents the number of days over which the return is calculated, with $n > 0$.

In case multiple events of the same type appear in different days, regarding the same asset, the return is averaged for the number of days, as follows:

$$R_i = \frac{\sum_{j=1}^{N} r_j}{N} \ , \tag{2}$$

where N is the number of days where events of this type occurred.

To correct the returns for the general market sentiment, we focus on excess returns. The excess return is calculated as the individual return of an asset that is achieved in excess of the market return, i.e., the return of the main index in which the asset is included:

$$a_i = r_i - r_i^I \ , \tag{3}$$

where r_i^I denotes the return of the index employed as benchmark.

When dealing with multiple events of a certain type appearing in different days and with excess returns, we correct these returns for the number of days:

$$A_i = \frac{\sum_{j=1}^{N} a_j}{N} \ , \tag{4}$$

where N is the number of days where events of this type occurred.

For the results presented in this section, the benchmark index used to compute excess returns is the FTSE350 index. An overview of the results is presented in Table 1. For each event we compute the absolute and excess returns for the day of the event, R_0 and A_0, as well as the returns following the event one, two, five, and ten days after the event is made public. For each of the events, we compute the percentage of events for which the direction of the return corresponds to the sign of the impact, i.e., positive returns in the case of positive impacts and negative returns in the case of negative impacts, and we denote this by d. Additionally, we compute the two-tailed t-test significance for the returns obtained for each of the events, and report this as p. The impacts reported in Table 1 have been manually determined by finance experts from Semlab. Last, three remarks are in place when interpreting these results:

- Multiple events may determine asset prices, and thus the returns, while not all these events could be captured through the news messages used for the analysis. However, we assume that the largest share of the reported returns is captured by the reported events.
- Reactions to events, in terms of price changes, may initiate before the event is public. By relying on the asset's closing price on the day previous to the event, we incorporate most of the anticipation preceding an event.
- When manually assessing an event's impact on stock prices (reported under the impact column in Table 1) the assumption is made that no other interactions, involving, for example, other events, have a significant influence on the price.

An initial inspection of Table 1 reveals that in nearly 90% of the event types, the direction of the R_0 returns corresponds with the sign of the impact assessed by experts. The two events where this is not the case are the *collaboration consideration* and *performance meets expectations* events. However, the expert impact associated with these events is only slightly positive, while the generated returns are slightly negative. Thus, based on the small number of events on which this impact is assessed, the assumptions listed in the previous paragraph, and the small difference between the expert impact and the generated returns, we consider the impact assigned by experts to be trustworthy in the absence of additional data. Last, the slightly negative returns are not significant at the 95% level.

When considering R_0, the event that generates the highest return is the *shares up* event, producing an average of 1.63%, backed up by the fact that 85% of this type of events generated a positive return. Presumably, not all events in this category generate a positive return due to the fact that, in some occasions, this event co-occurs with another event that generates a decrease in price that dominates the increase associated with the *shares up* event.

In the short run, i.e., when considering the R_0, R_1, and R_2 returns, we find three events for which the returns are both statistically significant, as well as showing the same direction as the impact determined by experts, in all three cases: *shares up*, *shares down*, and *rating up*.

In the long run, i.e., when considering the R_5 and R_{10} returns, we find more events for which both returns are statistically significant as well as being correctly captured by the manually determined impacts: *shares up*, *rating up*, *rating positive*, *profit up*, and *shares down*. From this we conclude that, for most events, the impacts are observable at longer time intervals after the event is reported.

When excess returns are considered, the short run exhibits four events for which the returns are significant at the 95% level and the direction of the return corresponds with the sign of the impact: *rating up*, *shares up*, *acquisition start*, and *shares down*. For the long run, the same is found for the events: *rating up*, *shares up*, *rating positive*, *price target raised*, *profit up*, and *shares down*.

TABLE 1
Average returns R_x and abnormal returns A_x for different time intervals of x days after an event.

Event	Impact	Freq.	R_0	d	p	R_1	d	p	R_2	d	p	R_5	d	p	R_{10}	d	p
SharesUp	2	756	1.63	85	0.00	1.65	80	0.00	1.53	73	0.00	1.74	72	0.00	2.27	69	0.00
RatingUp	2	36	1.55	83	0.00	1.82	72	0.00	1.89	75	0.00	2.52	69	0.00	2.31	72	0.01
CollabStart	2	13	1.06	46	0.32	1.73	62	0.18	1.95	62	0.14	1.95	77	0.10	1.30	62	0.20
RatingPositive	1	12	0.84	75	0.25	0.96	58	0.32	1.51	75	0.28	3.23	75	0.02	4.26	92	0.02
ProfitExceedsExp	3	18	0.74	50	0.39	1.99	61	0.11	1.72	61	0.19	2.87	56	0.08	2.61	61	0.12
AcquisitionStart	3	111	0.40	62	0.08	0.56	59	0.06	0.53	55	0.12	0.90	59	0.02	0.90	62	0.06
SalesUp	2	22	0.33	64	0.44	0.56	55	0.37	0.35	55	0.65	1.17	64	0.29	1.44	59	0.23
PriceTargetRaised	2	77	0.24	51	0.31	0.54	56	0.13	0.65	58	0.11	0.83	56	0.08	2.35	71	0.00
BusinessExpand	1	32	0.21	53	0.70	0.83	56	0.16	0.69	56	0.34	1.52	66	0.08	2.13	72	0.02
JointVenture	1	95	0.18	53	0.31	0.12	47	0.59	0.10	52	0.72	0.49	56	0.14	0.78	53	0.05
PerformExceedsExp	3	36	0.11	58	0.82	0.35	50	0.52	0.25	47	0.70	1.49	53	0.11	0.58	56	0.53
ProfitUp	2	110	0.08	50	0.80	0.22	50	0.53	0.26	52	0.54	1.25	56	0.03	1.73	64	0.01
CollabConsider	1	59	-0.05	49	0.81	-0.08	54	0.83	-0.17	49	0.69	0.19	59	0.67	0.11	59	0.85
PerformMeetsExp	1	25	-0.21	40	0.66	-0.11	48	0.79	0.31	56	0.57	0.55	64	0.46	0.75	56	0.42
ProfitDown	-2	27	-0.94	48	0.21	-0.52	52	0.49	0.05	48	0.95	0.33	52	0.78	0.83	52	0.56
RatingDown	-2	13	-0.96	54	0.15	-0.98	62	0.14	-1.32	69	0.11	-0.65	54	0.54	0.04	46	0.98
PriceTargetLowered	-2	26	-1.17	73	0.16	-1.44	77	0.12	-1.77	73	0.07	-1.50	65	0.14	-1.51	50	0.20
SharesDown	-2	630	-1.38	81	0.00	-1.49	71	0.00	-1.44	69	0.00	-1.20	62	0.00	-0.91	59	0.00
ProfitLessExp	-3	14	-2.52	64	0.06	-2.33	71	0.04	-2.04	71	0.08	-1.38	57	0.40	-1.29	64	0.35
		2,112		60			60			61			62			62	

Event	Impact	Freq.	A_0	d	p	A_1	d	p	A_2	d	p	A_5	d	p	A_{10}	d	p
RatingUp	2	36	1.49	81	0.00	1.88	72	0.00	1.89	72	0.00	2.46	69	0.00	2.02	69	0.03
SharesUp	2	756	1.32	80	0.00	1.31	74	0.00	1.26	68	0.00	1.59	68	0.00	1.58	64	0.00
CollabStart	2	13	1.02	46	0.26	1.62	69	0.15	1.76	62	0.12	1.74	77	0.10	0.83	62	0.23
ProfitExceedsExp	3	18	0.98	67	0.24	1.90	61	0.13	1.52	61	0.23	2.38	67	0.12	1.48	50	0.35
RatingPositive	1	12	0.83	67	0.26	1.21	67	0.17	1.69	75	0.11	2.88	75	0.03	3.35	67	0.03
AcquisitionStart	3	111	0.44	60	0.03	0.59	56	0.03	0.59	52	0.05	0.61	50	0.10	0.49	49	0.25
PerformExceedsExp	3	36	0.37	53	0.33	0.46	61	0.29	0.44	53	0.38	1.15	53	0.17	0.22	47	0.79
PriceTargetRaised	2	77	0.30	52	0.15	0.69	53	0.04	0.85	58	0.02	0.85	49	0.05	1.70	61	0.00
ProfitUp	2	110	0.19	46	0.48	0.43	55	0.15	0.45	55	0.17	1.21	55	0.01	1.44	56	0.01
SalesUp	2	22	0.16	59	0.69	0.39	59	0.47	0.26	55	0.72	1.04	55	0.27	0.94	59	0.35
BusinessExpand	1	32	0.15	38	0.74	0.66	56	0.23	0.46	50	0.46	1.17	53	0.10	1.11	53	0.18
JointVenture	1	95	0.14	49	0.36	0.18	54	0.39	0.13	51	0.58	0.13	49	0.65	-0.02	44	0.96
CollabConside	1	59	-0.14	44	0.48	-0.10	53	0.75	-0.20	49	0.57	-0.24	58	0.49	-0.63	41	0.16
PerformMeetsExp	1	25	-0.21	40	0.64	-0.32	40	0.47	-0.01	44	0.99	-0.27	48	0.67	-0.58	52	0.51
PriceTargetLowered	-2	26	-0.94	73	0.24	-1.36	85	0.16	-1.59	73	0.12	-2.24	73	0.04	-2.34	58	0.07
RatingDown	-2	13	-1.11	69	0.08	-0.91	69	0.13	-1.10	69	0.20	-0.40	46	0.71	-1.06	54	0.42
ProfitDown	-2	27	-1.14	63	0.14	-0.75	52	0.31	-0.19	52	0.82	0.22	63	0.83	0.12	63	0.93
SharesDown	-2	630	-1.18	80	0.00	-1.30	73	0.00	-1.30	73	0.00	-1.49	68	0.00	-1.46	64	0.00
ProfitLessExp	-3	14	-2.42	64	0.09	-2.42	71	0.04	-2.31	79	0.05	-1.36	71	0.30	-2.19	79	0.12
		2,112		59			62			61			60			57	

Event: financial event
Impact: predefined predicting factor for future price movement
Freq.: frequency of the event
R_x: average returns after x days
A_x: average abnormal returns after x days
d: percentage of returns which went into the right direction (positive if impact is positive)
p: two-tailed t-test significance at the 95% level

The linear relationship between the predefined event impact and the generated absolute returns is quantified by means of Person's correlation. For all time intervals in Table 2, we find a significant, positive correlation between the returns and impacts.

Additionally, we report the values for Pearson's correlation for the predefined impact and abnormal returns. Again, for all time horizons, we find strong and significant positive correlations. In the case of excess returns, the values for Pearson's correlation are higher than in the case of absolute returns, indicating that correcting for the index leads to results that are more in line with the expectations.

Based on the results, two conclusions can be drawn. First, the events that are selected and extracted from news messages can be employed in trading strategies, as in most cases these events provide the ability to generate positive returns. Second, the predefined impact associated with the extracted events is a good reflection of the impact of these events on stock prices, as apparent from the Pearson correlation test presented in this section.

4 Technical Trading

This section focuses on the technical trading indicators used in trading strategies generated through genetic programming. The indicators included in the study are: the simple moving average (SMA), the Bollinger band (BB), the exponential moving average (EMA), the rate of change (RoC), momentum (MOM), and

TABLE 2
Pearson's correlation between impact and returns R_x, and between impact and abnormal returns A_x.

R_x	r	p	A_x	r	p
R_0	0.844	0.000	A_0	0.878	0.000
R_1	0.851	0.000	A_1	0.863	0.000
R_2	0.804	0.000	A_2	0.827	0.000
R_5	0.789	0.000	A_5	0.753	0.000
R_{10}	0.662	0.002	A_{10}	0.712	0.001

R_x: average returns after x days
A_x: average abnormal returns after x days
r: Pearson's correlation between impact and R_x or A_x
p: two-tailed t-test significance at the 99% level

moving average convergence divergence (MACD). The choice for these indicators is based on their widespread use in technical trading [29].

4.1 Simple Moving Average

The SMA averages the last 20 days of the price of a stock [29], and is computed as:

$$M_i = \frac{\sum_{i=1}^{N} P_i}{N} , \tag{5}$$

where P_i represents the price on day i. The average is calculated over a fixed period of 20 days prior to the day for which the average is calculated, i.e., $N = 20$, which is standard for this indicator. A buy signal is generated when the price crosses the moving average in an upward movement, while a sell signal is generated when the price crosses the moving average in a downward movement.

4.2 Bollinger Bands

The Bollinger band is a technical indicator which creates two 'bands' around a moving average [29]. These bands are based on the standard deviation of the price. It is assumed that the price will move within these bands, around the moving average. If the volatility is high, the bands are wide and when there is little volatility the bands are narrow. The lower and upper Bollinger bands can be calculated as:

$$L = M - 2 \times \sigma_M , \tag{6}$$

$$U = M + 2 \times \sigma_M , \tag{7}$$

where σ_M stands for the volatility of moving average M. A buy signal is generated when the price is below the lower band, which is regarded as an oversold situation. A sell signal is generated at an overbought situation, when the price is above the upper band.

4.3 Exponential Moving Average

The exponential moving average (EMA) aims to identify trends by using a short and a long term average [29]. When the averages cross each other, it is the start of a new trend. The short term average is set at 5 days and the long term average at 20 days:

$$E_i = \frac{2}{N+1} \times (P_i - E_{i-1}) + E_{i-1} , \tag{8}$$

where P_i represents the price on day i, and N is the number of days. The initial EMA is calculated using the SMA, in our case for 5 and 20 days respectively starting from the first observation, as previously described. When the short term average crosses the long term average upwards, a buy signal is generated. A sell signal is generated when the short term average crosses the long term average downwards.

4.4 Rate of Change

The rate of change (RoC) is an indicator that calculates the difference between the closing price P_i of the current day i and the closing price P_{i-10} of 10 days earlier [29], according to the following equation:

$$C_i = \frac{P_i - P_{i-10}}{P_{i-10}} . \tag{9}$$

If the RoC starts decreasing above 0 (a peak was reached), a sell signal is generated. If it starts increasing below 0, a buy signal is generated.

4.5 Momentum

The momentum indicator uses exactly the same formula as the RoC. Instead of creating a buy signal after a peak, it creates a buy signal when the momentum crosses the 0 level upwards [29]. A sell signal is generated when the RoC crosses the 0 level downwards.

4.6 Moving Average Convergence Divergence

The moving average convergence divergence (MACD) is a technical indicator that subtracts two exponential averages from each other, namely the 12 and the 26 day exponential average [29]. The mathematical formula for the MACD is:

$$D_i = E[12]_i - E[26]_i . \tag{10}$$

A buy signal is generated when the MACD reaches the 0 level in an upward motion. A sell signal is generated when the MACD breaks through the 0 level in a downward motion.

4.7 Performance of Technical Trading Indicators

We now analyze the performance of the individual technical indicators when considered separately from any other indicators. In Table 3, we present the returns generated by each technical indicator over different time intervals. On the one hand, the focus of this table is on the buy signals generated by the indicators. The frequency shows how many buy signals are generated by the indicator. The returns show the average return surrounding a buy signal in the given time frame, e.g.,

TABLE 3
Returns for signals generated by technical indicators.

	Buy Signals						Sell signals					
Indicator	Freq.	R_0	R_1	R_2	R_5	R_{10}	Freq.	R_0	R_1	R_2	R_5	R_{10}
SMA(20)	1,663	**1.927**	**2.069**	**2.197**	**2.463**	**3.512**	1,737	**-1.888**	**-1.932**	**-1.923**	**-1.496**	**-0.805**
BB	2,014	-1.608	-1.374	-1.170	**0.110**	**0.180**	2,971	1.563	1.530	1.499	1.595	1.827
EMA(5,20)	870	**1.717**	**1.860**	**1.859**	**2.122**	**3.350**	922	**-1.662**	**-1.654**	**-1.586**	**-0.933**	**-0.488**
RoC(10)	4,387	-0.417	-0.290	-0.245	-0.053	**0.245**	2,937	0.723	0.637	0.953	1.564	2.370
MOM	1,988	**1.499**	**1.444**	**1.637**	**1.938**	**2.759**	2,049	**-1.310**	**-1.151**	**-1.248**	**-1.021**	**-0.629**
MACD(12,26)	667	**1.310**	**1.324**	**1.309**	**1.568**	**2.882**	581	**-1.276**	**-1.106**	**-1.162**	**-0.106**	0.317

Indicator: technical indicator generating buy and sell signals
Freq.: frequency of the signals
R_x: average returns after x days

R_{10} represents the 10 day return. The best performing technical indicator is the simple moving average, followed by the exponential moving average.

Table 3 additionally displays the performance of the technical indicators when sell signals are considered. Note that negative returns in this case are desirable, since higher magnitudes of a negative return indicate better performance of the sell signal (you would loose less money if you sell). Again, the two best performing technical indicators are the simple moving average and the exponential moving average.

5 A News-Based Trading Framework

Next, we introduce a framework for incorporating news in stock trading strategies. The framework assumes that events have been extracted from news messages and are available together with the date on which the events took place. Additionally, a predefined impact should be assigned to each event, allowing the news variable to be included in the trading strategies. For deriving the optimal trading strategies, we rely on genetic programming.

Genetic programming [30] is a technique where the potential solutions are represented as computer programs rather than numerical values encoded in some manner. Starting from a (usually randomly generated) initial population, genetic programs attempt to improve the fitness of individuals over successive generations through a process inspired by natural evolution. During this process, individuals are altered, usually based on their fitness values, by combining them with other individuals (crossover), or by slightly modifying some parts of the individual with a predefined probability (mutation). In this paper, genetic programming is used for finding optimal trading strategies based on technical indicators and news. Genetic programming has previously been used in the design of decision support systems, e.g., in [31], [32].

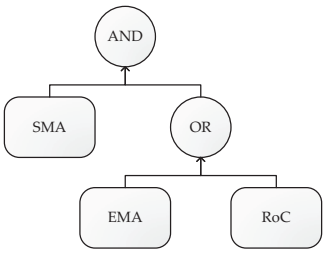

Fig. 2. Trading rule.

The trading strategies we determine take the form of trees that, when evaluated, return a Boolean value: true, when a trading signal is generated, or false, when no signal is generated and thus no action has to be taken. The trading strategies include at least one technical indicator or a news variable. Most often, the trading strategies include multiple variables, that may be either technical indicators or the news variable, connected by the logical operators 'and' and 'or'. An example of a trading strategy that may be generated is given in Fig. 2. This rule generates a trading signal when the simple moving average generates a trading signal simultaneously with at least one of the exponential moving average and rate of change indicators. The fitness of a trading strategy is computed based on the return that it generates on the dataset that we use. The returns are computed as indicated in Sect. 3.

The employed genetic programming algorithm for determining the optimal trading strategies is presented in Algorithm 1. We start from a random initial population of trees, and generate new populations of trading strategies by applying crossover and mutation on the population from the previous iteration. Crossover consists of selecting two trading strategies, and determining two random crossover points, i.e., one for each tree. Next, the subtrees generated under the crossover point are exchanged between the two trading strategies, thus resulting in two new rules that are added to the new population. Mutation only relates to the technical indicators included in a trading strategy, and consists of a slight change in the parameters of the randomly selected technical indicator, e.g., changing the number of days used by the simple moving average from 5 to 7. The stopping condition for the algorithm relates to the improvement in the best solution found, i.e., when the optimal solution cannot be improved in a number of generations, the algorithm stops.

We summarize the proposed framework in Fig. 3. As illustrated in the figure, the events are extracted

Algorithm 1 Genetic programming.

Require: $\alpha \geq 0$: minimum improvement
$\beta > 0$: maximum times of no improvement
$\gamma > 0$: population size
$0 < \rho \leq \gamma$: number of parents
$0 \leq \mu \leq 1$: mutation probability

π = generateRandomPopulation(γ)
$\sigma_{\text{old}} = -\infty$, σ_{new} = calcFitness(π), $b = 0$
while $b < \beta$ **do**
  addIndividual(π', getBest(π, σ_{new}))
  while $|\pi'| < |\pi|$ **do**
    θ = selectRandomParents(π, σ_{new}, ρ)
    ϑ = crossOver(θ)
    ϑ' = mutate(ϑ, μ)
    addIndividual(π', ϑ')
  end while
  $\pi = \pi'$, $\sigma_{\text{old}} = \sigma_{\text{new}}$, σ_{new} = calcFitness(π)
  if $\sigma_{\text{new}} - \sigma_{\text{old}} \leq \alpha$ **then**
    $b = b + 1$
  else if $b > 0$ **then**
    $b = 0$
  end if
end while
return π

from the news messages represented in free text format, and constitute the input to the algorithm. The historical price data constitutes an individual input to the search algorithm, used for computing the performance of the trading strategies, but is simultaneously used to derive the values for technical trading indicators, another input to the algorithm. Finally, the optimal trading strategies are determined through genetic programming.

6 Experiments and Results

In this section we provide an overview of the validation of our proposed framework for including news in stock trading strategies. First, we focus on the performance of the news variable taken individually, and then in combination with each of the technical indicators we consider. We then present the performance of optimal trading rules as determined through genetic programming, and discuss these results.

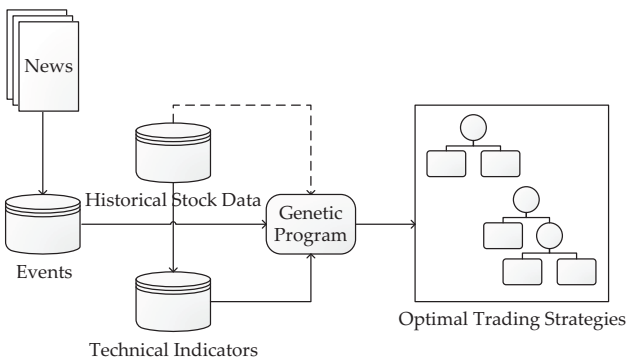

Fig. 3. News-based trading framework.

TABLE 4
Returns for signals generated by news.

Signal	Freq.	R_0	R_1	R_2	R_5	R_{10}
Buy	818	1.563	1.625	1.531	1.808	2.301
Sell	579	-1.578	-1.640	-1.606	-1.332	-1.061

Signal: type of signal generated by news
Freq.: frequency of the signal
R_x: average returns after x days

6.1 Performance of Individual Events

When trading strategies are built only by using the news variable, we generate the returns displayed in Table 4. Here, a buy signal is generated when events are encountered that are known to produce an R_0 of at least 0.5%, as shown in Table 1. Similarly, a sell signal is generated when the R_0 is below -0.5%. In Table 4 we present the results for buy and sell signals individually, for different time horizons.

A comparison with the results obtained by using trading strategies based only on technical indicators, as presented in Table 3, reveals that the news variable performs well. In the case of buy signals, the trading strategies based solely on the news variable are consistently outperformed by the simple moving average and the exponential moving average, but only slightly. In case of sell signals, the situation is similar, although the exponential moving average is not consistently outperforming the trading strategies based on the news variable.

The good overall performance obtained by using technical indicators and news, respectively, suggests that combining these indicators could result in better trading strategies. In the next section, we look at how each of the technical indicators performs when considered in combination with the news variable.

6.2 News and Technical Indicators

In this section we consider the performance of the individual technical indicators when trading strategies combine each of them with the news variable. Again, buy and sell signals are considered separately, and the returns are presented for different time horizons.

Table 5 presents the returns generated with buy signals when news and technical indicators are considered together. Positive returns generated through buy signals are displayed in bold, as these are considered desirable results. An overall conclusion is that when news and technical indicators are considered together, the generated returns are higher than when these variables are considered individually. Out of the six combinations, four consistently generate positive returns at all time horizons. The highest observed

TABLE 5
Returns for signals generated by news and technical indicators.

	Buy Signals						Sell signals					
Indicator	Freq.	R_0	R_1	R_2	R_5	R_{10}	Freq.	R_0	R_1	R_2	R_5	R_{10}
News & SMA(20)	108	**3.005**	**2.888**	**2.748**	**3.236**	**3.954**	93	**-3.177**	**-3.700**	**-3.849**	**-3.353**	**-3.166**
News & BB	18	-0.841	-0.587	**0.011**	**0.374**	-0.174	19	0.702	0.445	0.485	0.311	**-1.163**
News & EMA(5,20)	54	**3.284**	**3.081**	**3.013**	**3.433**	**4.246**	42	**-3.790**	**-3.894**	**-3.977**	**-3.529**	**-3.654**
News & RoC(10)	68	-0.098	**0.036**	-0.194	-0.060	**0.369**	28	0.211	**-0.414**	0.231	0.319	0.196
News & MOM	100	**3.058**	**2.857**	**2.976**	**3.339**	**3.874**	78	**-3.304**	**-3.748**	**-3.969**	**-3.417**	**-3.369**
News & MACD(12,26)	33	**3.566**	**3.610**	**3.563**	**4.040**	**5.371**	31	**-3.796**	**-4.217**	**-4.307**	**-4.555**	**-3.240**

Indicator: technical indicator generating buy and sell signals
Freq.: frequency of the signals
R_x: average returns after x days

return is generated by the combination of news and moving average convergence divergence, for time horizon R_{10}.

In Table 5 we additionally present the returns generated through sell signals when the individual technical indicators are considered together with news. Again, results are presented for different time horizons. The overall conclusion is that the combination of the two indicators generally outperforms the trading strategies based on the indicators taken separately. From the six combinations, four consistently generate desirable returns at all time horizons. Again, the highest return is achieved through the combination of news and moving average convergence divergence, but this time at time horizon R_5.

The results presented in this section allow us to conclude that combining individual technical indicators with the news variable for determining trading strategies enables higher returns than when technical indicators and the news variable are considered separately. We next move on to generating more complex trading strategies, that rely on multiple technical indicators, possibly in combination with the news variable, for generating trading signals.

6.3 Optimal Trading Strategies

The optimal trading strategies are determined through genetic programming, as outlined earlier. The initial population employed by the genetic program consists of 50 randomly generated trading rules. For our experiments, we let the algorithm run for 15 generations with a mutation rate of 0.5. For each new generation, all parents are selectable for crossover. Experiments are performed on both buy and sell signals, and take into consideration a selection of 2,000 data points from our dataset. Initial experiments have shown that these settings generate optimal results, while minimizing processing time. Strategies are generated for holding stocks for 1, 3, and 5 days.

In Table 6 we show the results obtained when the fitness of the generated trading rules is computed as the relative return after a number of days from the generation of a signal. The table displays the results of trading strategies making use of buy signals and sell signals.

All but one of the best performing buy rules when considering a one-day period include news as a relevant variable, with generated returns of 2.17% to

TABLE 6
Optimal strategies if stocks are held x days (FTSE350 dataset).

	Buy Signals			Sell signals		
x	Tree	Freq.	R_x	Tree	Freq.	R_x
1	SMA(27) & News	21	2.388	MOM(7) & MOM(3)	31	-2.333
	SMA(27) & SMA(15) | News	34	2.225	MOM(7) & SMA(28)	13	-2.315
	SMA(20) & MOM(5)	33	2.201	SMA(27) & News	23	-2.233
	SMA(27) & (SMA(17) | News)	41	2.177	MOM(7) & MOM(4)	29	-2.084
	(SMA(27) & SMA(21)) | (SMA(15) | News)	48	2.166	MOM(7) & News	28	-2.048
3	MOM(4) | News	34	2.636	SMA(24) & News	20	-2.578
	MOM(4) & SMA(27)	30	2.345	(MOM(4) & News) & MOM(5)	36	-2.464
	SMA(15) & SMA(16)	86	1.889	MOM(4) & News	36	-2.464
	SMA(17) | SMA(20)	120	1.627	MOM(7) & MOM(5)	43	-2.131
	News	196	1.220	SMA(24) & (MOM(5) & News)	24	-2.088
5	SMA(19) & News	28	2.246	MOM(7) & (SMA(24) | SMA(21))	22	-2.965
	SMA(22) & SMA(27)	50	2.081	MOM(7) & SMA(16)	38	-2.783
	SMA(18) & News	29	2.045	MOM(7) & News	28	-2.598
	SMA(24) & (MOM(6) | News)	34	1.919	News	147	-1.369
	MOM(3) | EMA(21,83)	180	1.570	(SMA(23) | News) & (SMA(24) | BB)	214	-1.368

x: number of days
Tree: generated trading strategy
Freq.: frequency of the signals
R_x: average returns after x days

2.39%. The simple moving average is included in all the rules, confirming the performance this indicator achieved when trading rules were considered that take into account only the individual technical indicators. When observing the top five trading strategies and their generated returns three days after the generation of a buy signal, we again note that the simple moving average is included in many trading strategies. Also, the news variable is part of the best performing trading rule, generating a return of 2.64%. However, when the time horizon consists of three days, the news variable is included less often in the optimal trading strategies, while momentum is of greater importance than before. For five-day rules, again, the simple moving average is included in many well performing trading strategies. News is included in the optimal trading strategy, as well as in two others, generating returns between 1.92% and 2.25%. Also, momentum and the exponential moving average can be found in the best trading strategies.

The table additionally shows the resulting top trading rules using the returns of sell signals and suggests that, as is the case when considering buy signals, news is an important indicator both on the short and on the long run, although for sell signals, news tends to be slightly less important for one-day sell strategies and more important for three- and five-day strategies. When observing one-day sell strategies, news is only included in two rules, whereas strategies for holding stocks a longer period include news more frequently. Average returns are comparable to the ones from the buy signals as well, yet it should be noted that momentum plays a more important role as a technical indicator for sell rules than for buy rules, and the simple moving average is used less frequently.

In order to validate the results, we perform the same experiments on another dataset, concerning all 500 companies listed under S&P500 at September 1st, 2012, covering the period between May 1st, 2010 and September 30th, 2010. News is collected through the Reuters news feed and stocks are scraped from Yahoo! Finance ticker data. The set has similar characteristics as the default dataset that is used throughout this paper, although there are more types of events, albeit with a less frequent occurrence. Also, average returns are generally higher for all time windows. Pruning is performed in the same way as done with our default dataset, resulting in 2,308 events after removing simultaneous and duplicate events, rare events, and events occurring on non-trading days.

Table 7 presents an overview of the generated trading rules based on this dataset. All parameter settings for the genetic programming algorithm have remained unchanged. For both buy and sell signals, the rules presented in the table confirm that news is often included in trading rules and hence is an important factor in these rules. A more detailed analysis of the FTSE350 and S&P500 rules is presented below.

When analyzing the generated rules for buy signals, we observe that in each of the best performing rule groups, news is included as a signal. For each of the evaluated horizons, the generated trading rules for S&P500 buy signals show many resemblances with those for the FTSE350 dataset in terms of the technical indicators used. In both datasets, many rules include news and/or make use of the simple moving average. Also, the number of times news is included in the best performing trading rules is approximately the same for both datasets for each of the evaluated horizons. The returns on the other hand are slightly higher for the S&P500 dataset.

For sell signals, the role of news in trading rules for S&P500 equities is visible, yet it is less prominent. For each of the evaluated horizons, news is included

TABLE 7
Optimal strategies if stocks are held x days (S&P500 dataset).

	Buy Signals			Sell signals		
x	Tree	Freq.	R_x	Tree	Freq.	R_x
1	SMA(27)	27	2.599	SMA(22) & SMA(14)	48	-3.118
	News | (SMA(26) | News)	115	2.328	SMA(28) & SMA(22)	51	-3.095
	SMA(18) | News	142	2.236	SMA(22) & MOM(5)	43	-2.965
	SMA(18) | MACD(12,26)	150	2.190	(EMA(24,94) | SMA(23)) | News	133	-2.316
	SMA(26) | (MOM(5) | News)	264	2.095	News | (SMA(24) | EMA(24,94))	132	-2.285
3	MOM(4) & SMA(15)	51	3.349	SMA(16) | EMA(16,136)	126	-2.469
	EMA(22,105) | SMA(28)	98	2.743	EMA(25,145) | SMA(16)	125	-2.406
	SMA(28) | News	108	2.711	SMA(16) | News	136	-2.231
	(News | (SMA(22) | SMA(18))) | SMA(20)	127	2.671	MOM(4) | SMA(16)	305	-1.891
	SMA(20)	117	2.669	News | MOM(3)	282	-1.612
5	News | SMA(23)	123	2.877	SMA(25) | SMA(15)	183	-2.434
	SMA(22) | News	127	2.806	EMA(14,137) | SMA(25)	106	-2.375
	(SMA(20) | News) | News	133	2.624	SMA(15) | News	144	-2.042
	SMA(25) | SMA(15)	118	2.613	News | SMA(27)	117	-1.931
	(EMA(18,105) | News) | SMA(27)	118	2.582	EMA(19,111) | MOM(5)	212	-1.621

x: number of days
Tree: generated trading strategy
Freq.: frequency of the signals
R_x: average returns after x days

in the best performing trading rules, yet not in the top rules. This could be explained by the fact that the frequency of events that trigger sell signals in this dataset when compared to our default dataset is lower. The latter observation underlines a key issue. When there is enough breaking news, information can play a crucial role in trading. However, when more news is published on less significant events, it is a less reliable trading indicator. Moreover, the signal scarcity is reflected in the operands used in the generated trading rules. While on our FTSE350 dataset, often the '&' operand is used for news inputs, on the S&P500 dataset the '|' operand is more frequently used. This implies that news is more often a strengthening signal for the technical indicators in the FTSE350 dataset compared to the S&P500 dataset. Another observation that can be made is that the momentum technical indicator is used less often, in favor of the simple moving average indicator, causing the compositional differences between rules for buy and sell signals to be smaller now. Furthermore, generated returns for the various resulting trading rules are different from the ones created for the FTSE350 dataset, which is caused by the fact that the S&P500 dataset covers different equities and a different time span.

7 Practical Considerations

When applying the proposed techniques to a new domain, one should consider the various steps of transforming news into events, converting events to signals, and generating rules based on signals. Of crucial importance are having at one's disposal a domain expert who is able to define concepts and (impacts of) events, a small training set for determining signals, and computing power for learning trading rules. Subsequently, the learned rules can be deployed in real-time applications for the considered domain.

Transforming news into signals for trading algorithms requires the extraction of features and events from text. Our proposed natural language processing approach, which makes use of part-of-speech tagging, lemmatizing, and – most importantly - ontology concept identification, can be employed for extracting domain-specific events from text. Our ViewerPro-based implementation is flexible in that the concepts and events of interest can be specified by the user. Hence, extending the news processing and event recognition approach to other domains can be achieved with minimal additional effort. In case of new concepts or events, expert knowledge can be employed for defining concept specifications and event impacts. Subsequently, the event-associated returns can be computed based on a training set in order to determine event signals that comprise the input of our subsequent trading rule learning algorithm.

In practice, the results of this study can be directly applied to trading algorithms by generating news-driven signals, indicating whether the user should buy, sell, or hold a specific stock. These signals can be generated as proposed in this paper. Alternatively, more complex signals can be obtained, which can be based not only on news, but also on other (numerical) inputs. When events extracted from news are turned into trading signals, one is able to devise trading rules that make use of various signals, not only stemming from news, but also originating from various other (numerical) inputs such as moving averages. As we successfully empirically demonstrated for two domains related to FTSE350 and S&P500 companies, by following our genetic programming approach, one could learn well-performing trading rules for a specific domain and for a range of horizons by constructing a set of rules that use signals from news or other numerical inputs, and by iteratively mutating and crossovering the best performing rules (i.e., the ones that generate the highest returns) in a subsequent processing step. In order to create good trading rules for a specific domain or scenario, the algorithm parameters such as mutation and crossover rates need to be adjusted to the scenario at hand, by means of the procedure explained below.

In order to apply our approach on a new market (e.g., NASDAQ100 companies), one should learn parameters based on a representative training set for the specific domain. For this, simple learning approaches such as a hill climbing procedure or more advanced approaches like genetic algorithms or other meta-heuristics can be used. In all cases, optimal parameter values are determined while maximizing returns. Additionally, if performance is a limitation, execution time should also be minimized. Parameters to be optimized are the initial population size (i.e., the number of rules), the number of generations (iterations), the mutation rate (i.e., the fraction or percentage of rules that are mutated), and crossover (i.e., the percentage of parent rules that are eligible for creating new offspring).

Although the main purpose of this paper is not to find the best trading rules by using news, the presented results can be applied to (trading) algorithms that are used in daily practice. For this purpose, additional aspects need to be taken into account, as for example the transaction costs and the rule creation speed. Existing trading algorithms (many proprietary) could benefit from our approach by additionally employing the news signals as shown in our proposed framework. As our framework is customizable by means of its various parameters, it can be used for other stock markets than the ones considered here, providing for a general methodology of including the news component in a trading algorithm.

8 Conclusions and Future Work

We presented a framework for incorporating news into stock trading strategies. The trading strategies

that we consider may include (in addition to the news variable) any number of technical trading indicators. The news variable is quantified based on the events extracted from the text of news messages and the assignment of an expert-defined impact to each of these events. Our results indicate that the assigned impacts correlate well with the returns generated by these events when tested against real data based on FTSE350 equities.

The selected technical indicators are also tested, and the individual performance of each indicator is reported. Additionally, combinations of individual technical indicators and the news variable are investigated. The results indicate that adding the news variable to each of the indicators generates higher returns than when each of the variables is considered alone. This suggests that considering the news variable indeed can lead to higher returns, thus making it worthwhile to employ trading rules that, next to technical indicators, make use of the events that are relevant for a certain company.

Last, a genetic program is used to discover complex trading rules based on technical indicators and news-based signals. For this purpose, we consider three time horizons when computing the fitness of the trading strategies based on the generated returns, namely one, three, and five days after the generation of a buy or sell signal. We conclude that, in many cases, news is a relevant variable for trading rules, and its inclusion in trading strategies leads to higher returns than when this variable is not considered. Experiments on a contrastive dataset containing data on S&P500 equities confirm the previous observations, and hence underline the importance of taking into account news as an additional input for trading rules. Based on the positive returns of the generated rules, we also conclude that the proposed framework is appropriate for including news in technical trading strategies.

Our results indicate that the inclusion of news into stock trading strategies can be achieved by extracting the events from the text of the news messages and associating an impact with these events (based on stock price variations for an event). This impact can later be used in the derivation of optimal trading strategies, where the news variable, consisting of the predefined impact, is used next to technical indicators. Returning to the two hypotheses stated in the introduction, namely that news will be included in the optimal trading strategies if news is a relevant variable and that these trading rules should generate positive returns, we conclude that the news variable has been quantified in a meaningful way, confirming our first hypothesis that news would be included in the optimal trading strategies. Additionally, all trading strategies that include news events generate a positive return, thus confirming our second hypothesis.

Future work will focus on including more indicators, technical or non-technical in nature, in the variable pool from which trading strategies are generated. Additionally, a more fine-grained analysis of the news messages, e.g., identification of event-related information such as the involved actors, should provide more information for generating trading strategies. Last, considering the interaction between events occurring within the same day, or within finer-grained time intervals, will provide a deeper understanding of the way that news impact stock prices and may lead to more profitable trading strategies.

ACKNOWLEDGMENTS

The authors are partially supported by the NWO Physical Sciences Free Competition project 612.001.009: Financial Events Recognition in News for Algorithmic Trading (FERNAT) and by the Dutch national program COMMIT.

REFERENCES

[1] J. Borsje, F. Hogenboom, and F. Frasincar, "Semi-automatic financial events discovery based on lexico-semantic patterns," *International Journal of Web Engineering and Technology*, vol. 6, no. 2, pp. 115–140, 2010.

[2] W. IJntema, J. Sangers, F. Hogenboom, and F. Frasincar, "A lexico-semantic pattern language for learning ontology instances from text," *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 15, no. 1, pp. 37–50, 2012.

[3] S. R. Das and M. Y. Chen, "Yahoo! for Amazon: Sentiment extraction from small talk on the web," *Management Science*, vol. 53, no. 9, pp. 1375–1388, 2007.

[4] M. Cecchini, H. Aytug, G. J. Koehler, and P. Pathak, "Making words work: Using financial text as a predictor of financial events," *Decision Support Systems*, vol. 50, no. 1, pp. 64–175, 2010.

[5] P. C. Tetlock, "Giving content to investor sentiment: The role of media in the stock market," *Journal of Finance*, vol. 62, no. 3, pp. 1139–1168, 2007.

[6] ——, "More than words: quantifying language to measure firms' fundamentals," *Journal of Finance*, vol. 63, no. 3, pp. 1437–1467, 2008.

[7] W. Leigh, R. Purvis, and J. M. Ragusa, "Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support," *Decision Support Systems*, vol. 32, no. 4, pp. 361–377, 2002.

[8] K. Mehta and S. Bhattacharyya, "Adequacy of training data for evolutionary mining of trading rules," *Decision Support Systems*, vol. 37, no. 4, pp. 461–474, 2004.

[9] M.-A. Mittermayer and G. F. Knolmayer, "Text mining systems for market response to news: a survey," Institute of Information Systems University of Bern, Tech. Rep., 2006, From: http://www.ie.iwi.unibe.ch/publikationen/berichte/resource/WP-184.

[10] M. L. Mitchell and J. H. Mulherin, "The impact of public information on the stock market," *Journal of Finance*, vol. 49, no. 3, pp. 923–950, 1994.

[11] W. S. Chan, "Stock price reaction to news and no-news: drift and reversal after headlines," *Journal of Financial Economics*, vol. 70, no. 2, pp. 223–260, 2003.

[12] S. T. Kim, J. C. Lin, and M. B. Slovin, "Market structure, informed trading, and analyst' recommendations," *Journal of Financial and Quantitative Analysis*, vol. 32, no. 4, pp. 507–524, 1997.

[13] S. G. Ewalds, M. B. J. Schauten, and O. W. Steenbeek, "De informatiewaarde van kwartaalcijfers," *Maandblad voor Accountancy en Bedrijfseconomie*, no. 7/8, pp. 333–341, 2000.

[14] J. B. Warner, R. L. Watts, and K. H. Wruck, "Stock prices and top management changes," *Journal of Financial Economics*, vol. 20, no. 1, pp. 461–492, 1988.
[15] K. A. Bonnier and R. F. Bruner, "An analysis of stock price reaction to management change in distressed firms," *Journal of Accounting and Economics*, vol. 11, no. 1, pp. 95–106, 1989.
[16] A. J. Keown and J. M. Pinkerton, "Merger announcements and insider trading activity: an empirical investigation," *Journal of Finance*, vol. 36, no. 4, pp. 855–869, 1981.
[17] R. J. Rosen, "Merger momentum and investor sentiment: the stock market reaction to merger announcements," *Journal of Business*, vol. 79, no. 2, pp. 987–1017, 2006.
[18] D. L. Ikenberry and S. Ramnath, "Underreaction to self-selected news events: the case of stock splits," *Review of Financial Studies*, vol. 15, no. 2, pp. 489–526, 2002.
[19] R. Michaely, R. H. Thaler, and K. L. Womack, "Price reactions to dividend initiations and omissions: overreaction or drift," *Journal of Finance*, vol. 50, no. 2, pp. 573–608, 1995.
[20] X. F. Zhang, "Information uncertainty and stock returns," *Journal of Finance*, vol. 61, no. 1, pp. 105–137, 2006.
[21] J. van Bommel, "Rumors," *Journal of Finance*, vol. 58, no. 4, pp. 1499–1520, 2003.
[22] D. L. McGuinness and F. van Harmelen, "OWL Web Ontology Language: W3C Recommendation 10 February 2004," 2004, From: http://www.w3.org/TR/owl-features/.
[23] V. Milea, F. Frasincar, and U. Kaymak, "Knowledge Engineering in a Temporal Semantic Web Context," in *Eighth International Conference on Web Engineering (ICWE 2008)*. IEEE Computer Society Press, 2008, pp. 65–74.
[24] F. Frasincar, V. Milea, and U. Kaymak, "tOWL: Integrating Time in OWL," in *Semantic Web Information Management: A Model-Based Perspective*, R. De Virgilio, F. Giunchiglia, and L. Tanca, Eds. Springer, 2010, pp. 225–246.
[25] V. Milea, F. Frasincar, K. Kaymak, and G. Houben, "Temporal optimizations and temporal cardinality in the tOWL language," *International Journal of Web Engineering and Technology*, vol. 7, no. 1, pp. 45–64, 2012.
[26] V. Milea, F. Frasincar, and K. Kaymak, "tOWL: a temporal web ontology language," *IEEE Transactions on Systems, Man and Cybernetics, Part B, Cybernetics*, vol. 42, no. 1, pp. 268–281, 2012.
[27] F. Allen and R. Karjalainen, "Using genetic algorithms to find technical trading rules," *Journal of Economics*, vol. 51, no. 2, pp. 245–271, 1999.
[28] F. Hogenboom, M. de Winter, M. Jansen, A. Hogenboom, F. Frasincar, and U. Kaymak, "Event-based historical value-at-risk," in *IEEE Computational Intelligence for Financial Engineering & Economics 2012 (CIFEr 2012)*. IEEE Computer Society Press, 2012, to appear.
[29] S. B. Achelis, *Technical Analysis from A to Z*. McGraw-Hill, 2000.
[30] J. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, Massachusetts, USA: MIT Press, 1992.
[31] H. Zhao, "A multi-objective genetic programming approach to developing Pareto optimal decision trees," *Decision Support Systems*, vol. 43, no. 3, pp. 809–826, 2007.
[32] W. Fan, P. Pathak, and L. Wallace, "Nonlinear ranking function representations in genetic programming-based ranking discovery for personalized search," *Decision Support Systems*, vol. 42, no. 3, pp. 1338–1349, 2006.

Wijnand Nuij obtained his MSc degree in Informatics & Economics from Erasmus University Rotterdam, the Netherlands, in 2009. Currently, he is a Knowledge Engineer at SemLab, an SME active in the Information Technology and Services industry. Part of his work is the management of the ViewerPro software for automatic news analysis with respect to a financial trader portfolio. Wijnand Nuij is also a guest lecturer on algorithmic trading at the Erasmus University Rotterdam.

Viorel Milea obtained the MSc degree in Informatics & Economics from Erasmus University Rotterdam, the Netherlands, in 2006. In 2013, he received his PhD degree at the Erasmus University Rotterdam. His research focuses on employing Semantic Web technologies for enhancing the current state-of-the-art in automated trading with a focus on processing information contained in economic news messages and assessing its impact on stock prices. Other research interests cover areas such as Semantic Web theory and applications, intelligent systems in finance, and nature-inspired classification and optimization techniques.

Frederik Hogenboom obtained cum laude the MSc degree in Economics & Informatics from the Erasmus University Rotterdam, the Netherlands, in 2009, specializing in computational economics. During his bachelor and master programmes, he published research mainly focused on the Semantic Web and learning agents. Currently, he is active within the multidisciplinary field of business intelligence and continues his research in a PhD track at the Erasmus University Rotterdam. His PhD research focuses on ways to employ financial event discovery in emerging news for algorithmic trading, hereby combining techniques from various disciplines, amongst which Semantic Web, text mining, artificial intelligence, machine learning, linguistics, and finance. Other research interests are related to applications of computer science in economic environments, agent-based systems, and applications of the Semantic Web.

Flavius Frasincar obtained the MSc degree in computer science from Politehnica University Bucharest, Romania, in 1998. In 2000, he received the PDEng degree in software engineering from Eindhoven University of Technology, the Netherlands. He got the PhD degree in computer science from the Eindhoven University of Technology, in 2005. Since 2005, he is assistant professor in information systems at Erasmus University Rotterdam, the Netherlands. He has published in numerous conferences and journals in the areas of databases, Web information systems, personalization, and the Semantic Web. He is a member of the editorial board of the International Journal of Web Engineering and Technology.

Uzay Kaymak received the MSc degree in electrical engineering, the Degree of Chartered Designer in information technology, and the PhD degree in control engineering from the Delft University of Technology, Delft, the Netherlands, in 1992, 1995, and 1998, respectively. From 1997 to 2000, he was a Reservoir Engineer with Shell International Exploration and Production. He holds the chair of information systems in the healthcare at the School of Industrial Engineering, Eindhoven University of Technology, the Netherlands. Prof. Kaymak has co-authored more than 200 academic publications in the fields of intelligent decision support systems, computational intelligence, data mining, and computational modeling methods. He is an associate editor of IEEE Transactions on Fuzzy Systems and is a member of the editorial board of several journals.