Quantitative Finance
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rquf20

Can a corporate network and news sentiment improve portfolio optimization using the Black-Litterman model?

Germán G. Creamer[a]
[a] Howe School, Stevens Institute of Technology, 1 Castle Point on the Hudson, Hoboken, NJ, 07030USA.
Published online: 02 Jun 2015.

To cite this article: Germán G. Creamer (2015) Can a corporate network and news sentiment improve portfolio optimization using the Black-Litterman model?, Quantitative Finance, 15:8, 1405-1416, DOI: 10.1080/14697688.2015.1039865

To link to this article: http://dx.doi.org/10.1080/14697688.2015.1039865

Quantitative Finance, 2015
Vol. 15, No. 8, 1405–1416, http://dx.doi.org/10.1080/14697688.2015.1039865

Routledge
Taylor & Francis Group

Can a corporate network and news sentiment improve portfolio optimization using the Black–Litterman model?

GERMÁN G. CREAMER*

Howe School, Stevens Institute of Technology, 1 Castle Point on the Hudson, Hoboken, NJ 07030, USA

(*Received 16 June 2013; accepted 23 March 2015*)

The Black–Litterman (BL) model for portfolio optimization combines investors' expectations with the Markowitz framework. The BL model is designed for investors with private information or knowledge of market behaviour. In this paper, I propose a method where investors' expectations are based on either news sentiment using high-frequency data or on a combination of accounting variables; financial analysts' recommendations, and corporate social network indicators with quarterly data. The results show promise when compared to a market portfolio. I also provide recommendations for trading strategies using the results of this BL model.

Keywords: Link mining; Social network; Machine learning; Boosting; Text analysis; Portfolio optimization; Black–Litterman model

JEL Classification: C44, C58, C63, G11

1. Introduction

Contemporary investment literature is significantly influenced by Markowitz (1952, 1959)'s portfolio optimization approach, which suggests an optimal allocation of assets that maximizes expected return and minimizes volatility. Unfortunately, this mean-variance portfolio optimization process may lead to the selection of a few top assets and is very sensitive to small changes in inputs. It is also based on past price history and investors cannot formally add their own knowledge of the market. As a reaction to these limitations, Black and Litterman (1990) proposed a mean-variance portfolio optimization model that included investors' expectations. This methodology, the Black–Litterman (BL) model, creates views that represent investors' market expectations with different confidence levels and uses these views as inputs for the selection of the optimal portfolio.

There are two different approaches which have not been deeply explored in the investment literature: the application of link mining and text analysis to solve finance problems.

Link mining, the first approach, is a set of techniques that uses different types of networks and their indicators to forecast or model a linked domain. Link mining has had several applications (Senator 2005) to different areas such as money laundering (Kirkland *et al.* 1999), telephone fraud detection (Fawcett and Provost 1999), crime detection (Sparrow 1991), and surveillance of the NASDAQ and other markets (Kirkland *et al.* 1999, Goldberg *et al.* 2003). One of the most important business applications of link mining is in the area of viral marketing or network-based marketing (Domingos and Richardson 2001, Hill *et al.* 2006, Leskovec *et al.* 2006, Richardson and Domingos 2006), and more recently, in finance. Adamic *et al.* (2009) have applied network analysis to quantify the flow of information through financial markets. Creamer and Stolfo (2009) have applied a link mining algorithm called CorpInterlock to integrate the metrics of an extended corporate interlock (social network of directors and financial analysts) with corporate fundamental variables and analysts' predictions (consensus). CorpInterlock used these metrics to forecast the trend of the cumulative abnormal return and earnings surprise of US companies.

The second approach, the application of text analysis for financial forecasting, has become more important in the recent years due to the large amount of unstructured data available in large corporate and public databases. A particular application has been the study of financial news (Devitt and Ahmad 2007, Haider and Mehrotra 2011, Xie *et al.* 2013), tweets (Bar-Haim *et al.* 2011), blogs (Ruiz *et al.* 2012) or boards (Chua *et al.* 2009) for sentiment analysis. In the area of portfolio optimization, Seo *et al.* (2004) and Decker *et al.* (1996) describe a multi-agent portfolio management system that automatically classifies financial news. Thomas (2003) combines news classification with

*Email: gcreamer@stevens.edu

technical analysis indicators in order to generate new trading rules. Lavrenko *et al.* (2000) describe a system that recommends news stories that can affect market behaviour. Wuthrich *et al.* (1998) and Cho *et al.* (1999) weigh keywords based on their occurrences to predict the direction of major stock indices. Text classification and retrieval applied to finance is still an area underexplored in the literature. However, several investment banks and hedge funds are quickly developing systems to automatically incorporate the impact of daily news into their trading systems. These studies suggest that the sentiment of the news articles is predictive of the sentiment of investors, which in turn affects their trading behaviour and subsequently the prices of securities.

In this paper, I test two mechanisms to simulate the investors' view of the BL model: (1) the quarterly return forecast generated by PortInterlock, which is a variation of the CorpInterlock algorithm, and (2) news sentiment using high-frequency and daily data. This methodology may help investors to incorporate qualitative and quantitative factors into their investment decisions without the intervention of financial management experts.

2. Methods

The simulation of the investors' view requires the combination of the BL portfolio approach with machine learning, link mining and automated text analysis algorithms as described in this section.

2.1. The BL model

The BL model is one of the most extended tactical allocation models used in the investment industry. The BL model calculates the posterior excess return using a mean-variance optimization model and the investors' view. Since the introduction of the BL model (Black and Litterman 1990), many authors have proposed several modifications. Khrishnan and Mains (2005) extend the BL model to include a factor uncorrelated with the market; Becker and Gurtler (2010) and Beach and Orlov (2007) substitute the investors' views with analysts' dividend forecast and GARCH derived views, respectively.

I follow Black and Litterman (1992, 1999) in describing the BL model. Additional useful references about the BL model are Idzorek (2009) and Walters (2009).

The excess returns of n assets over the risk free rate R_f are normally distributed and are represented by the n-vector μ. Using a Bayesian framework, the prior distribution of excess returns is $\mu \sim N(\Pi, \Sigma)$ where Π is a n-vector of implied equilibrium excess return and Σ is the nxn variance covariance matrix of excess return.

Σ can be obtained by the historical excess return and the equilibrium excess return $\Pi = \lambda \Sigma w$ is the solution to the following unconstrained return maximization problem:

$$\max_{w} w'\Pi - \frac{\lambda w' \Sigma w}{2} \tag{1}$$

where λ is the risk aversion parameter. The vector of optimal portfolio weights can be derived from the Π formula:

$$w = (\lambda \Sigma)^{-1} \Pi \tag{2}$$

In equilibrium, the market portfolio w_{mkt} derived from the capital asset pricing model (CAPM) should be the same as the mean-variance optimal portfolio w, so the prior expected excess return should be the equilibrium expected excess return:

$$\Pi = \lambda \Sigma w_{mkt} \tag{3}$$

where $w_{mkt} = \frac{M_i}{\sum_i M_i}$ and M_i is the market capitalization value of asset i.

The main innovation of the BL model is that the investor may specify k absolute or relative scenarios or 'views' about linear combinations of the expected excess return of assets. The views are independent of each other and are also independent of the CAPM. They are represented as:

$$P\mu = Q - \epsilon \tag{4}$$

P is a kxn matrix where each row represents a view. Absolute views have weights for the assets that will outperform their expected excess returns and their total sum is one; relative views assign positive and negative weights to assets that over- or underperform, respectively, and their total sum is zero. Q is a k-vector that represents the expected excess return of each view, τ is a k-vector that represents the confidence indicator of each view and $\epsilon \sim N(0, \Omega)$ is an error term normally distributed that represents the uncertainty of the views where Ω is the kxk diagonal covariance matrix of error terms of the views.

The posterior distribution of excess returns $\hat{\mu}$ combines the prior excess return Π and the investors' views P:

$$\hat{\mu} = N\left(\left[(\tau\Sigma)^{-1} + P'\Omega^{-1}P\right]^{-1}\left[(\tau\Sigma)^{-1}\Pi + P'\Omega^{-1}P\right], \left[(\tau\Sigma)^{-1}\Pi + P'\Omega^{-1}P\right]^{-1}\right) \tag{5}$$

An alternative expression of the expected excess return is:

$$\hat{\mu} = \Pi + \tau\Sigma P'\left[P\tau\Sigma P' + \Omega\right]^{-1}[Q - P\Pi] \tag{6}$$

and the vector of optimal portfolio weights on the unconstrained efficient frontier using the posterior distribution is:

$$\hat{w} = (\lambda\Sigma)^{-1}\hat{\mu} \tag{7}$$

2.2. Learning algorithms

2.2.1. Classification and regression trees (CART). CART, a very popular decision tree algorithm proposed by Breiman *et al.* (1984), builds a binary decision tree following a top–down approach where the best feature with the best threshold for separating the data in two parts according to a test such as the information gain is introduced as the root node where its branches are the values of this feature. This process is repeated successively with the descendants of each node creating two new nodes until there is no further information gain or any other stopping rule is satisfied. An alternative implementation is to grow the tree as much as possible and then prune each node that most improves accuracy. At that point, a leaf node is included with the most common value of the target attribute.

The information gain ($\Delta E(S, A)$) for the introduction of the target feature A and the sample of training observation S is defined as:

$F_0(x) \equiv 0$
for $t = 1 \ldots T$
$w_i^t = \frac{1}{1+e^{y_i F_{t-1}(x_i)}}$
Get h_t from *weak learner*
$\alpha_t = \frac{1}{2} \ln \left(\frac{\sum_{i:h_t(x_i)=1, y_i=1} w_i^t}{\sum_{i:h_t(x_i)=1, y_i=-1} w_i^t} \right)$
$F_{t+1} = F_t + \alpha_t h_t$

Figure 1. The LogitBoost algorithm Friedman *et al.* (2000). y_i is the binary label to be predicted, x_i corresponds to the features of an instance i, w_i^t is the weight of instance i at time t, and h_t and $F_t(x)$ are the prediction rule and the prediction score at time t, respectively.

$$\Delta E(S, A) \doteq E(S) - \sum_{v \in values(A)} \frac{S_v}{S} E(S_v)$$

and entropy impurity is:

$$E(S) \doteq - \sum_{i=1}^{c} p_i \log_2 p_i$$

where p_i is the proportion of observations S that belongs to class i and c represents the different values of the target feature

Considering the popularity of CART in the financial service industry, I use it as the baseline algorithm.

2.2.2. Boosting. AdaBoost is a machine learning algorithm invented by Freund and Schapire (1997) that classifies its outputs by applying a simple learning algorithm (weak learner) to several iterations of the training set in which the misclassified observations receive more weight. Freund and Mason (1999) proposed a decision tree learning algorithm called an *alternating decision tree* (ADT). In this algorithm, boosting is used to obtain the decision rules and to combine them using a weighted majority vote.

Friedman *et al.* (2000), followed by Collins *et al.* (2004) suggested a modification of AdaBoost, called LogitBoost. LogitBoost can be interpreted as an algorithm for step-wise logistic regression. This modified version of AdaBoost—known as LogitBoost—assumes that labels $y_i's$ were stochastically generated as a function of $x_i's$. Then, it includes $F_{t-1}(x_i)$ in the logistic function to calculate the probability of y_i, and the exponent of the logistic function becomes the weight of the training examples. Figure 1 describes LogitBoost.

2.2.3. PortInterlock: a link mining algorithm. CorpInterlock is a link mining algorithm proposed by Creamer and Stolfo (2009) to build a bipartite social network: one partition includes members of board of directors and another partition consists of financial analysts representing the companies they cover. This social network is converted into a one-mode network in which the vertices represent the companies and the edges represent the number of directors and analysts that every pair of companies have in common. This is the extended corporate interlock. The basic corporate interlock is calculated in the same way using only directors. The algorithm selects the largest strongly connected component of a social network and ranks its vertices using a group of investment variables presented in appendix 1 and a group of social network statistics obtained from the basic or extended corporate interlock. Finally, the algorithm predicts the trend of a financial time series using a machine learning algorithm such as boosting. I propose the PortInterlock algorithm, an extension of CorpInterlock, to be used for portfolio optimization. This algorithm uses the financial time series predictions trend as the view of the investors and prior asset excess returns to define the optimal portfolio weights (figure 2).

Forecasting earnings surprise:

I used the definition of earnings surprise or forecast error proposed by Dhar and Chou (2001):

$$FE \doteq \frac{\text{CONSENSUS}_q - \text{EPS}_q}{|\text{CONSENSUS}_q| + |\text{EPS}_q|}$$

where CONSENSUS_q is the mean of earnings estimate by financial analysts for quarter q and EPS_q is the actual earnings per share for quarter q. FE is a normalized variable with values between -1 and 1. Additionally, when CONSENSUS_q is close to zero and EPS_q is not, then the denominator will not be close to zero.

The increasing importance of organizational and corporate governance issues in the stock market suggests that the integration of indicators from the corporate interlock with more traditional economic indicators may improve the forecast of FE and CAR.

The following indicators obtained by the PortInterlock algorithm capture the power relationship between directors and financial analysts:

(i) Degree centrality: directors and analysts of a company characterized by a high degree or degree centrality coefficient are connected through several companies.
(ii) Closeness centrality: directors and analysts of a company characterized by a high closeness centrality coefficient are connected through several companies that are linked through short paths.
(iii) Betweenness centrality: directors and analysts of a reference company characterized by a high betweenness centrality coefficient are connected through several companies. Additionally, the reference company mentioned above has a central role because it lies between several other companies, and no other company lies between this reference company and the rest of the companies.
(iv) Clustering coefficient: directors and analysts of a company characterized by a high clustering coefficient are probably as connected amongst themselves as is possible through several companies.

Each of the measures above shows a different perspective of the directors–analysts relationship. Hence, I could include them as features in a decision system to forecast FE and CAR. Because the importance of these features (combined with a group of financial variables to predict FE) may change significantly in different periods of time, I decided to use boosting, specifically LogitBoost, as the learning algorithm. Boosting is well known for its feature selection capability, its error bound proofs (Freund and Schapire 1997), its interpretability and its capacity to combine continuous and discrete variables. Creamer and Freund (2004, 2005, 2007) have already applied boosting to forecast equity prices and corporate performance

Input: Two disjoint nonempty sets V_{11} and V_{12}, a matrix ER of historical excess returns of each asset i, a financial time series Y to be predicted, the covariance matrix Ω of error terms of the views of investors, the vector τ that represents the confidence indicator of each view, the risk factor λ, a vector M with market capitalization values of each asset i, and additional exogenous variables.

1. Build a bipartite graph $G_1(V_1, E_1)$ in which its vertex set V_1 is partitioned into two disjoint sets V_{11} and V_{12}, such that every edge in E_1 links a vertex in V_{11} and a vertex in V_{12}.
2. Build a one-mode graph $G_2(V_2, E_2)$ in which there exists an edge between v_i and v_j: v_i, $v_j \in V_2$ if and only if v_i and v_j share at least a vertex $u_i \in V_{12}$. The value of the edge is equal to the total number of objects in V_{12} that they have in common.
3. Calculate the largest strongly connected component of G_2 and label it $G_3(V_3, E_3)$.
4. Calculate the adjacency matrix A and geodesic distance matrix D for G_3. a_{ij} and d_{ij} are the elements of A and D, respectively.
5. For each vertex $v_i \in V_3$ calculate the following social network indicators:
   - Degree centrality: $deg(v_i) = \sum_j a_{ij}$
   - Closeness centrality (normalized): $C_c(v_i) \doteq \frac{n-1}{\sum_j d_{ij}}$
   - Betweenness centrality: $B_c(v_i) = \sum_i \sum_j \frac{g_{kij}}{g_{kj}}$, where g_{kij} is the number of geodesic paths between vertices k and j that include vertex i, and g_{kj} is the number of geodesic paths between k and j.
   - Clustering coefficient: $CC_i = \frac{2|\{e_{ij}\}|}{deg(v_i)(deg(v_i)-1)} : v_j \in N_i,\ e_{ij} \in E$
   - Normalized clustering coefficient: $CC'_i = \frac{deg(v_i)}{MaxDeg} CC_i$, in which MaxDeg is the maximum degree of vertex in a network
6. Merge social network indicators with any other relevant set of variables for the population under study, such as analysts' forecasts and economic variables, and generate test and training samples.
7. Run a machine learning algorithm with above test and training samples to predict trends of Y.
8. Define a matrix P in which each row k is the multiplication of the confidence of the prediction and the prediction of the trends of Y for each asset. P represents the absolute view of the investors.
9. Obtain Q as a k-vector that represents the expected excess return of each asset or each view k, Σ as the variance covariance matrix of ER, and $\Pi = \lambda \Sigma w_{mkt}$ as the equilibrium expected excess return where $w_{mkt} = \frac{M_i}{\sum_i M_i}$
10. Optimize the portfolio using the Black Litterman model where the expected excess return is:
$\hat{\mu} = \Pi + \tau\Sigma P'[P\tau\Sigma P' + \Omega]^{-1}[Q - P\Pi]$
and the vector of optimal portfolio weights is $\hat{w} = (\lambda\Sigma)^{-1}\hat{\mu}$.
Output: Optimal portfolio weights (w).

Figure 2. The PortInterlock algorithm.

showing that LogitBoost performs significantly better than logistic regression, the baseline algorithm. Dhar and Chou (2001) have also compared tree-induction algorithms, neural networks, naive Bayesian learning and genetic algorithms to classify the earnings surprise before announcement.

2.3. Automated text analysis

Automated text analysis is typically used in finance and accounting research to solve text classification problems such as sentiment analysis using some type of supervised learning algorithms with features generated by the frequency of a series of n-grams or of keywords of a dictionary. Semantic analysis is also another important area of research in text analysis although social scientists have not given it as much attention as other methods. The text analysis methods used in this paper are the following:

2.3.1. Bag-of-words (BoW). The prevailing method of presenting text is BoW due to its robustness and simplicity (Hagenau *et al*. 2012). BoW defines for each document a set of weights given by the word or term frequency (tf_t) and/or by the term frequency-inverse document frequency ($tf-idf_t$) of each term t. $tf - idf_t = \log\frac{N}{df_t}$, N is the number of documents in the collection and df_t is the document frequency of the term t.

The main problem with BoW is that it does not identify the target that sentiment refers to. For example, 'X beats Y' is positive for X, but is negative for Y. BoW cannot provide such information. Since different words in various domains may present distinct sentiments, it is also impossible for BoW to identify the causality among the variables studied without semantic information. Since BoW uses every word as a feature, a large data-set generates a very large feature set and much irrelevant information. Therefore, an appropriate method to extract useful information is still critical to calculate sentiment (Forman 2003, Schumaker and Chen 2009, Hagenau *et al*. 2012).

2.3.2. Semantic frames. Frame semantic parsing of a document refers to the automated task of finding semantic targets, disambiguating their semantic frames that refer to particular events and identifying their frame elements. The theory of frame semantics (Fillmore 1982) has motivated the development of ontologies such as the FrameNet lexicon (Baker *et al*. 1998) that acts as a repository of semantic frames and their frame elements. Das *et al*. (2010) proposed a rule-based system called SEMAFOR for target identification. SEMAFOR uses a latent-variable log linear model to identify semantic frames and a probabilistic model to identify their frame elements. In this research, we used SEMAFOR to identify semantic frames with their targets and frame elements of every news.

2.3.3. Latent Dirichlet allocation. The topic model methodology (Steyvers and Griffiths 2007) discovers common topics among a series of documents. This approach assumes that documents are a mixture of topics, where a topic is based on a probability distribution over words. To evaluate a new document, topics are chosen according to their distribution, and keywords are associated with specific topics. By inverting this process, it is possible to infer the set of topics used to generate the documents. The latent Dirichlet allocation (LDA) is an accepted topic model methodology for capturing the latent structure of a large set of documents. LDA simply supposes that the topic distribution follows a Dirichlet prior (Blei *et al*. 2003). In this paper, this approach clusters topics across large data-sets of news.

The application of these unsupervised methods to business problems is still very limited. Aral *et al*. (2011) use LDA to extract common topics among 2397 stock recommendations, Creamer *et al*. (2013) apply LDA to extract common topics in a corporate network and use it to forecast return, Bao and Datta (2014) use an extended version of LDA topic model to evaluate the effect of risk disclosures in 10-K forms on the risk perception of investors and Xie *et al*. (2013) test several NLP methods such as BoW, LDA and semantic frames for stock price prediction.

3. Simulations

I conducted two groups of simulations to evaluate the integration of the BL model with different automated decision support systems and two different time horizons. The first group of experiments uses the output of the PortInterlock algorithm with US quarterly data to simulate the investors' view of the BL model, while the second group of experiments uses sentiment analysis calculated with text mining methods to optimize a portfolio based on high-frequency and daily data.

3.1. US market and the PortInterlock algorithm: quarterly data

The asset price and return series are restricted to the US stock market. They are from the Center for Research in Security Prices (CRSP), the accounting variables from COMPUSTAT,† the list of financial analysts and earnings forecast or consensus from IBES, and the annual list of directors for the period 1996–2005 is from the Investor Responsibility Research Center. The number of companies under study changes every year. The minimum and maximum number of companies included in my study is 3043 for 2005 and 4215 for 1998.

I implemented the PortInterlock algorithm (figure 2) with the software Pajek (de Nooy *et al*. 2005) to obtain the basic (social network of directors) and extended corporate interlock. I computed the investment signals as described in appendix 1 and the social network statistics introduced in the previous section of the basic and extended corporate interlock. Most of the fundamental and accounting variables used are well known in the finance literature and Jegadeesh *et al*. (2004) demonstrated that these variables are good predictors of cross-sectional returns. I merged the accounting information, analysts' predictions (consensus) and social networks statistics using quarterly data and selected the last quarter available for

†COMPUSTAT is an accounting database managed by Standard & Poor's.

every year. I forecasted the trend of FE and CAR. CAR is calculated using the cumulative abnormal return of the month following the earnings announcement. Every instance has the label '1' if the trend was positive and '−1' otherwise. CAR is calculated as the return of a specific asset minus the value weighted average return of all assets in its risk-level portfolio according to CRSP. FE is based on the predictions of the analysts available 20 days before the earnings announcement as fund managers may suggest (Dhar and Chou 2001). Fund managers take a position, short or long,† a certain number of days before the earnings announcement and, according to their strategy, they will liquidate the position a given number of days after the earnings announcement. Investors profit when the market moves in the direction expected and above a certain threshold, even though the market movement might not be in the exact amount forecasted.

I restricted my analysis to trading strategies using FE because the prediction of FE (test error of 19.09%) outperformed the prediction of CAR (test error of 47.56%). According to Creamer and Stolfo (2009), the long-only portfolio is the most profitable strategy when it is compared with a long-short, a long-short for the most precise decile, and a long-only strategy when analysts predict that earnings will be larger than consensus. Based on these results, the weights of the long-only portfolio multiplied by the confidence of the prediction are used as the investors' views of the BL model. As this model is the result of the PortInterlock algorithm, I call this model the BL-PortInterlock portfolio. This portfolio is compared against a market portfolio in which the weight of each asset is based on its market capitalization.

3.2. *European market and news: high-frequency and daily data*

This component includes two steps. The first step is to obtain news sentiment using high-frequency data, and in the second step, I use the news sentiment as the investors' view to rebalance the BL portfolio.

In the first step, I selected a stratified sample of 1000 news stories associated with STOXX 50 companies from the TR News Archive for the year 2005. This index represents the 50 most important European companies by level of capitalization. I used 743 news stories after eliminating news without proper prices; incomplete news, news that contained market reports and news summaries that included information from many companies simultaneously. I matched the timestamp of the news with the timestamp of the most relevant associated assets' prices. The news stories are labelled as positive or negative according to the asset return's direction and 0 otherwise after one, five and 15 min, and one, two and three hours. I associated every news story to a particular company using the Reuters Instrument Codes (RIC) based on the first main sentence of the story, which in most cases was the first RIC in the field Related_RICS of TR News Archive. However, I verified that the news was about the selected RIC. Whenever a news article first mentioned the company of an external analyst or a manager's opinion or shared information about a particular company, I selected the target company instead of the company of the analyst. I extracted the text of every article from the TR News Archive field Take_Text and I substituted any missing data of this field with the corresponding header (HEADLINE_ALERT_TEXT).

After eliminating stop words, I used the text analysis methods introduced above to extract the following quantitative features from the news and classify them according to the asset return direction using AdaBoost (see section 2.2.2) as the main learning algorithm:

(i) Bag-of-words (BoW): tf and $tf\text{-}idf$ of 1-, 2- and 3-grams for all words.
(ii) Part of speech (POS): tf and $tf\text{-}idf$ of 1-, 2- and 3-grams for all words, verbs only, adverbs only, adjectives only and nouns only.
(iii) Semantic frames (SF): tf and $tf\text{-}idf$ of 1-, 2- and 3-grams of semantic frames, targets and frame elements.
(iv) Latent Dirichlet analysis (LDA): topics generated.

I compared the results of AdaBoost with CART (see section 2.2.1), considering that CART is widely known in the financial services industry. I selected the year 2005 as this was the first year that I had a complete sample of news, and it also overlaps with the last year of the sample used for the BL-PortInterlock portfolio. I used 70% (520) of the observations to train the forecasting model and tested on the remaining 30% (223) observations.

In the second step, I restricted the high-frequency test sample to those news that were directly associated with a STOXX 50 company that have at least two news present in the test sample. As a result of this restriction, I used 125 news of 27 companies associated to the STOXX 50 index. I used the news sentiments for the period that showed the best prediction (3 h) as the investors' view of the BL model to optimize the European 'long only' portfolio to be consistent with the first group of simulations. If the news has a negative, neutral or positive sentiment, I assign 0, 1, or 2 or more, respectively, to the cell that represents the company that generates the news in the investors' view. In this paper, this portfolio is called BL-News.

I obtained the excess log return of every selected company after subtracting to the log return, the rate of the lowest euro area government bond whose rating is triple A (Beta 2) as provided by the European Central Bank. I used daily prices from 3 January to 7 September 2005 to calculate historical excess log returns and the variance covariance matrix. I tested the model using the complete period from 8 September to 30 December 2005 where the portfolio is rebalanced every time that at least one news arrives. During the test period, three trading days is the maximum period without relevant news about the portfolio. This period is acceptable considering that according to Creamer *et al.* (2013), news of the European STOXX 50 have a lagged impact of at least seven days.

I calculate the daily log return using the close prices of the day when the news arrives and the close prices of the next day. This calculation underestimates the immediate effect of the news, while our analysis with the high-frequency data indicates that the longer horizon (3 h) has a stronger return effect than the 1 min horizon, so the news impact is still very important.

†Long or short positions refer to buy a specific asset or to sell a borrowed asset based on the expectation that price of the asset will increase or decrease respectively.

I compare the different versions of the portfolio with the equally weighted portfolio, with the STOXX 50 index and with the market portfolio in which the weight of each asset is based on its market capitalization. The performance of the market portfolio and the STOXX 50 index is very similar, although they are not exactly the same. This difference is because the market capitalization portfolio includes 27 major European stocks that were part of the STOXX 50 index when this information was collected. However, the STOXX 50 index includes several other companies that are not part of the tests conducted in this paper. For this reason, the main benchmark for this study is the market portfolio.

4. Results

4.1. The BL model and the BL-PortInterlock portfolio

Table 1 compares the result of several views based on a portfolio completely generated by social networks and fundamental indicators (BL-PortInterlock) with the CorpInterlock portfolio, an equally weighted portfolio and the market portfolio. The CorpInterlock portfolio shows the largest Sharpe ratio (6.56), while the market portfolio has an annual Sharpe ratio of 1.415. Therefore, when the confidence in the investors' view based on CorpInterlock is very large ($\tau = 1$), the BL-PortInterlock portfolio has the highest Sharpe ratio and it approximates the results of the CorpInterlock portfolio. Likewise, when the confidence in the investors' view decreases (τ), the Sharpe ratio deteriorates. This can be explained because when τ is zero, the result of the BL model is the market portfolio as indicated by equation 6.

The *t*-test mean difference between each scenario and the market portfolio is significant for most of the cases. This test deteriorates as well as the Sharpe ratio when the covariance of the error term of the investors' view increases ($\omega = 0.01$).

Figure 3 indicates that the inclusion of social network and accounting indicators generates a portfolio (BL-PortInterlock) with a higher level of expected return than the equally weighted and the market portfolio as input. The inclusion of corporate social network indicators might capture interactions among directors and financial analysts that improve the prediction of earnings surprise. This effect, combined with the predictive capacity of selected accounting indicators, explains why a portfolio with a social network perspective outperforms the market portfolio.

Table 1. Annual Sharpe ratio by portfolio type, by the covariance of the error term of the investors' views (Ω) and by the confidence indicator of the investors' views (τ). BL-PortInterlock is the portfolio based on the PortInterlock algorithm.

Portolio/views	$\Omega = 0.0001$	$\Omega = 0.001$	$\Omega = 0.01$
BL-PortInterlock, CI, $\tau = 1$	4.782**	4.402**	0.022
BL-PortInterlock, CI, $\tau = 0.75$	4.767**	4.281**	1.433
BL-PortInterlock, CI, $\tau = 0.5$	4.736**	4.063**	1.863
BL-PortInterlock, CI, $\tau = 0.25$	4.646**	3.509**	1.791*
BL-PortInterlock, CI, $\tau = 0.01$	0.022	1.604*	1.437*
BL-PortInterlock, CI, $\tau = 0.001$	1.604*	1.437*	1.418*
CorpInterlock	6.563**		
Equally weighted	2.840*		
Market capitalization	1.415		

*,**: 5 and 1% significance level of *t*-test mean difference between each scenario and the market portfolio.

4.2. The BL model, news sentiment and the BL-News portfolio

Tables 2 and 3 indicate that AdaBoost using the features of BoW-POS-SF-LDA shows the best performance to forecast asset return in comparison to other text analysis methods according to the F-score and the test error. In both cases, the performance improves using the largest 3 h horizon. Additionally, AdaBoost outperforms CART with a 95% confidence level when both algorithms are trained with the features of BoW-POS-SF-LDA (see table 4).

Using the news sentiment generated by AdaBoost with the features of the combined text analysis method BoW-POS-SF-

Table 2. F-score of asset return forecast using AdaBoost by text analysis methods and time horizon in minutes (min) and hours (h).

	1 min	5 min	15 min	1 h	2 h	3 h	Mean
BoW-POS-SF-LDA	0.43	0.35	0.44	0.48	0.46	0.50	0.44
BoW-POS-SF	0.38	0.34	0.45	0.47	0.47	0.50	0.43
LDA	0.33	0.39	0.42	0.41	0.47	0.46	0.41
BoW-POS	0.37	0.36	0.39	0.46	0.48	0.41	0.41
BoW-SF	0.30	0.39	0.43	0.39	0.46	0.51	0.41
POS-SF	0.35	0.31	0.37	0.41	0.52	0.49	0.41
SF	0.30	0.37	0.42	0.41	0.43	0.50	0.41
POS	0.33	0.34	0.36	0.46	0.45	0.48	0.41
BoW	0.35	0.37	0.39	0.42	0.44	0.45	0.40
Mean	0.35	0.36	0.41	0.43	0.46	0.48	0.42

Table 3. Test error of asset return forecast using AdaBoost by text analysis methods and time horizon in minutes (min) and hours (h).

	1 min	5 min	15 min	1 h	2 h	3 h	Mean
BoW-POS-SF-LDA	0.57	0.65	0.55	0.49	0.53	0.48	0.55
BoW-POS-SF	0.62	0.65	0.55	0.50	0.51	0.48	0.55
LDA	0.62	0.58	0.55	0.57	0.48	0.50	0.55
BoW-POS	0.62	0.62	0.54	0.51	0.50	0.52	0.55
BoW-SF	0.70	0.60	0.57	0.59	0.50	0.47	0.57
POS-SF	0.64	0.70	0.63	0.59	0.49	0.50	0.59
SF	0.69	0.62	0.57	0.56	0.54	0.48	0.58
POS	0.66	0.66	0.64	0.54	0.54	0.51	0.59
BoW	0.64	0.59	0.60	0.54	0.55	0.54	0.58
Mean	0.64	0.62	0.57	0.54	0.52	0.50	0.56

Table 4. F-score of asset return forecast using AdaBoost and CART with the combined text analysis method BoW-POS-SF-LDA by time horizon in minutes (min) and hours (h).

	1 min	5 min	15 min	1 h	2 h	3 h	Mean
AdaBoost	0.43	0.35	0.44	0.48	0.46	0.50	0.44*
CART	0.27	0.36	0.43	0.44	0.37	0.48	0.39
Mean	0.35	0.36	0.44	0.46	0.41	0.49	0.42

*: 5% significance level (*p*-value) of *t*-test mean difference between AdaBoost and CART.

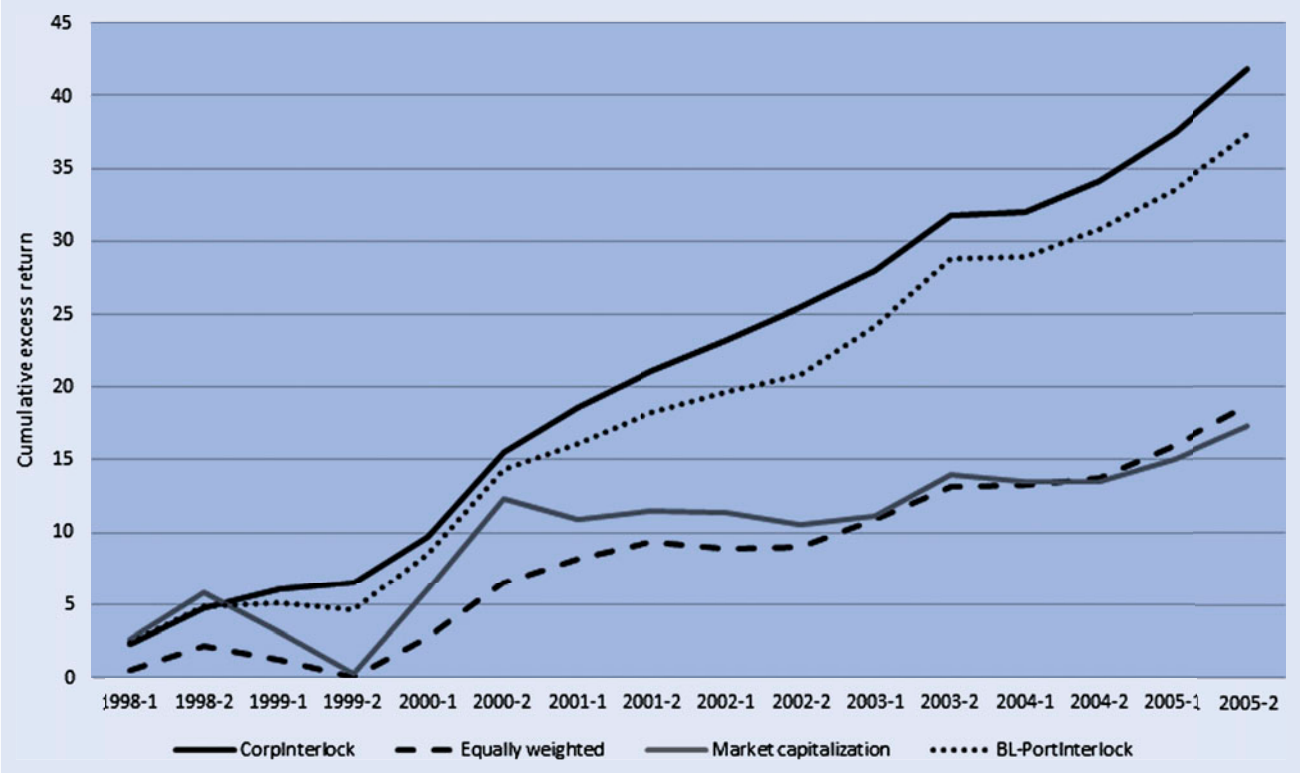


Figure 3. Cumulative excess log return of a BL portfolio based on the BL-PortInterlock portfolio, the CorpInterlock portfolio, an equally weighted portfolio and the market capitalization portfolio.

Table 5. Annual Sharpe ratio by portfolio type, the covariance of the error term of the investors' views (Ω) and the confidence indicator of a particular view (τ). BL-News is the news sentiment-driven BL portfolio.

Portolio/views	$\Omega = 0.0001$	$\Omega = 0.001$	$\Omega = 0.01$	$\Omega = 0.1$
BL-News, CI, $\tau = 1$	8.892**	8.818**	8.806**	8.805**
BL-News, CI, $\tau = 0.75$	8.870**	8.813**	8.805**	8.805**
BL-News, CI, $\tau = 0.5$	8.842**	8.809**	8.805**	8.804**
BL-News, CI, $\tau = 0.25$	8.817**	8.806**	8.805**	8.804**
BL-News, CI, $\tau = 0.1$	8.807**	8.805**	8.804**	8.804**
BL-News, CI, $\tau = 0.01$	8.804**	8.804**	8.804**	8.804**
Equally weighted	8.194**			
Market capitalization	4.461			
STOXX 50	5.272			

**: 1% significance level of *t*-test mean difference between each scenario and the market capitalization portfolio.

LDA as the investors' sentiment of the BL model, the scenario with large confidence on news sentiment ($\tau = 1$) and with low covariance of error term of the investors' views ($\Omega = 0.0001$) leads to the largest average Sharpe ratio (table 5). The different versions of the BL-News portfolio outperform the market portfolio.

The Sharpe ratio of the BL-News portfolio declines when the confidence on investors' views (τ) decreases or the covariance of the error term of the views (Ω) increases, although in a smaller proportion than in the case of the PortInterlock algorithm. Figure 5(a) shows that BL portfolios with low covariance of the error term ($\Omega = 0.001$) and high confidence ($\tau = 1$) of investors' view have different prior and posterior distributions of excess return. Therefore, if the investors' views are consistent with the future market conditions, the performance of these portfolios should outperform the market portfolio. However, when the confidence ($\tau = 0.1$) of investors' view is very low (figure 5(b)), then the prior and posterior probability are practically identical. In this particular case, the BL portfolio becomes the market portfolio as indicated by equation 6.

5. Discussion: dynamic optimization and transaction costs

The BL-PortInterlock portfolio is rebalanced only twice a year using quarterly data; however, its cumulative excess return is almost twice the return of the market portfolio. Considering the large income difference and the very infrequent adjustment of the BL-PortInterlock portfolio, the effect of transaction costs is very small and should not have any major impact on the long-term performance of the portfolio.

The main question is whether the same conclusion holds for the BL-News portfolio. Considering that the market portfolio used as a benchmark also has important transaction costs because it is rebalanced daily, and the cumulative excess log return of the BL-News portfolio almost duplicates the excess return of the market portfolio (figure 4), the difference of cumulative excess return between these portfolios is large enough to absorb the transactions costs. As expected, figure 4 indicates that the final cumulative excess returns of the STOXX 50 index and the market portfolio are very similar, although they differ in several time periods due to variations in their asset composition as explained in section 3.2.

The equally weighted portfolio shows a similar cumulative excess return as the market portfolio but has a higher Sharpe ratio in the two cases explored in this paper. The main reason for this is that the portfolio is not rebalanced. As a result, it has low volatility and minimum transaction costs, although it is not

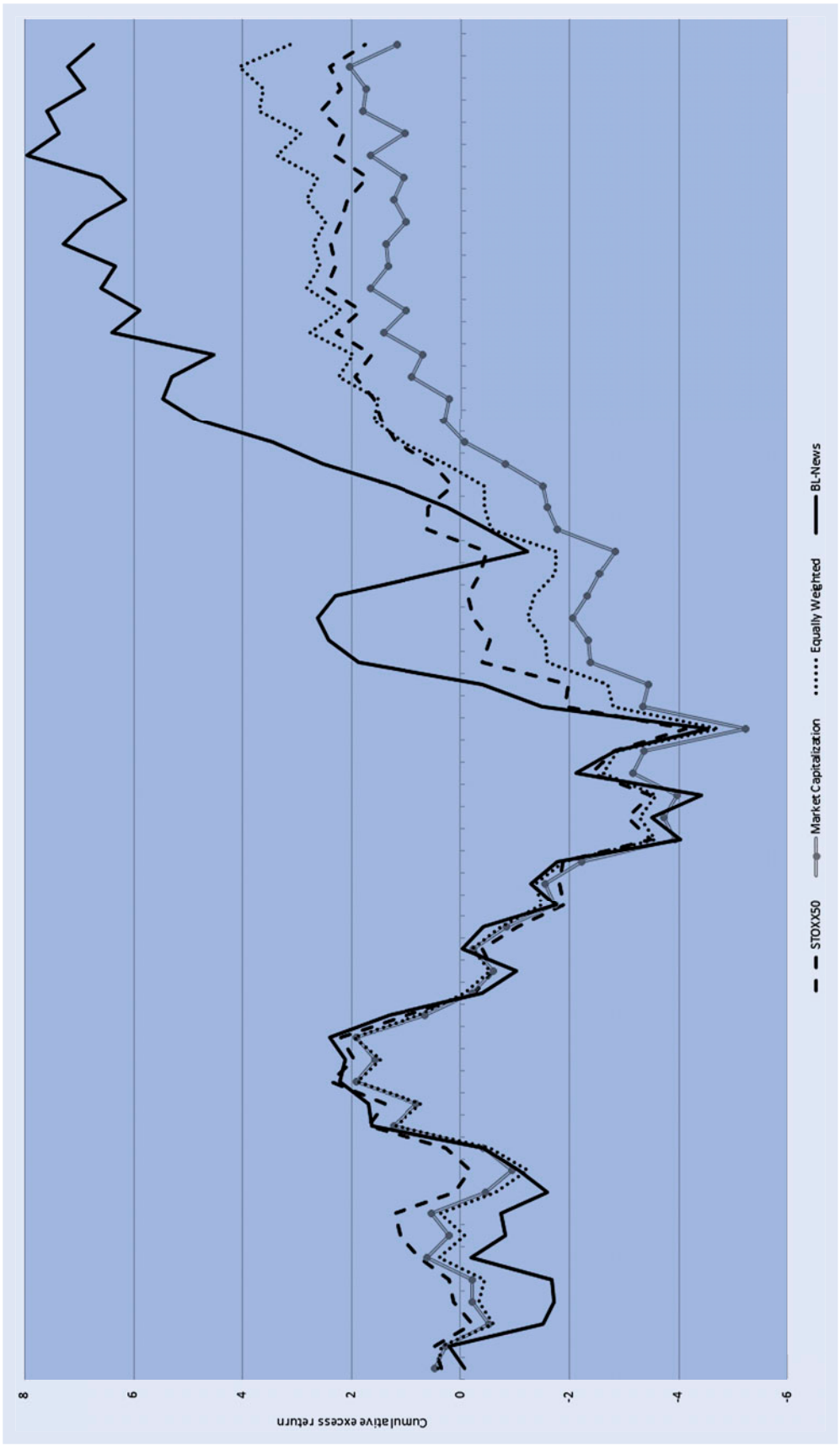
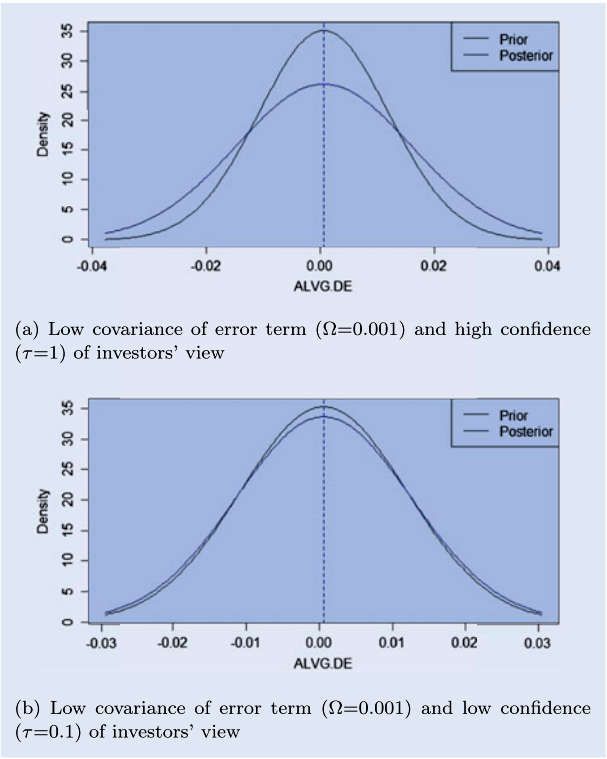


Figure 4. Cumulative excess log return of a news sentiment-driven BL portfolio (BL-News), an equally weighted portfolio, the market capitalization portfolio and the STOXX 50 index. The BL-News portfolio assumes a low covariance of error term ($\Omega = 0.001$) and high confidence ($\tau = 1$) of the investors' view.



(a) Low covariance of error term (Ω=0.001) and high confidence (τ=1) of investors' view

(b) Low covariance of error term (Ω=0.001) and low confidence (τ=0.1) of investors' view

Figure 5. Prior and posterior distribution of a sample asset's return according to the BL model.

as capable of adapting to the change of the market conditions as the BL portfolio using news or social network indicators. This explains why the Sharpe ratios of the BL-PortInterlock and BL-News portfolios are higher than the Sharpe ratio of the equally weighted portfolio. In addition, the cumulative excess returns of the BL portfolios are almost twice than the same return of the equally weighted portfolio. Therefore, from the point of view of investors, the extra effort required to use either social network indicators or news sentiment for portfolio optimization leads to an important improvement of excess returns.

6. Conclusions

This paper shows that a modified BL model, which includes a forecast based on quarterly data of social networks and fundamental indicators as investors' view, outperforms the market portfolio. This research also demonstrates that news sentiment has an important high-frequency effect on return and they can be used as a proxy for investors' view on a BL portfolio. The simulations indicate that a news sentiment-driven portfolio outperforms the market portfolio and the market index.

In conclusion, the investors' subjective views of the BL model can be substituted or enriched by forecasts based on the optimal combination of social networks, news sentiment and accounting indicators. This research can be extended using the combination of these different aspects to forecast asset return.

Acknowledgements

The author thanks Ionut Florescu, Maria Christina Mariani, Frederi G. Viens, two anonymous referees, and participants of the Financial Mathematics session of the American Mathematical Society meeting 2010, the Workshop on Information in Networks (WIN)-NYU 2010, the IEEE CEC 2011 Workshop on Agent-Based Economics and Finance, the Sixth Rutgers-Stevens Workshop Optimization of Stochastic Systems 2011, and the Eastern Economics Association meeting 2012 for their valuable comments. The author also thanks Patrick Jardine for proof-reading the article. This work was supported by the Howe School Alliance for Technology Management.

Disclosure statement

No potential conflict of interest was reported by the author.

References

Adamic, L., Brunetti, C., Harris, J. and Kirilenko, A., Information flow in trading networks. Paper presented at 1st Workshop on Information in Networks, NYU, New York, NY, 2009.

Aral, S., Ipeirotis, P.G. and Taylor, S., Content and context: Identifying the impact of qualitative information on consumer choice. In *Proceedings of the 32nd International Conference on Information Systems*, 2011 (AIS: Shanghai).

Baker, C.F., Fillmore, C.J. and Lowe, J.B., The Berkeley framenet project. In *Proceedings of the 17th International Conference on Computational Linguistics*, pp. 86–90, 1998 (Montreal, QC).

Bao, Y. and Datta, A., Simultaneously discovering and quantifying risk types from textual risk disclosures. *Manage. Sci.*, 2014, **60**, 1371–1391.

Bar-Haim, R., Dinur, E., Feldman, R., Fresko, M. and Goldstein, G., Identifying and following expert investors in stock microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1310–1319, 2011 (Association for Computational Linguistics: Edinburgh, Scotland, UK).

Beach, S. and Orlov, A., An application of the Black–Litterman model with EGARCH-M derived views for international portfolio management. *Financ. Markets Portfolio Manage.*, 2007, **21**, 147–166.

Becker, F. and Gurtler, M., Quantitative forecast model for the application of the Black–Litterman approach. 2009 Finance International Meeting AFFI – EUROFIDAI, Paris, 2010.

Black, F. and Litterman, R., *Asset Allocation: Combining Investor Views With Market Equilibrium. Fixed Income Research*, 1990 (Goldman Sachs: New York).

Black, F. and Litterman, R., Global portfolio optimization. *Financ. Anal. J.*, 1992, **48**, 28–43.

Black, F. and Litterman, R., *The Intuition Behind Black–Litterman Model Portfolios. Investment Management Research*, 1999 (Goldman Sachs: New York).

Blei, D.M., Ng, A.Y. and Jordan, M.I., Latent Dirichlet allocation. *J. Machine Learn. Res.*, 2003, **3**, 993–1022.

Borgatti, S.P. and Everett, M., A graph-theoretic perspective on centrality. *Soc. Netw.*, 2006, **28**, 466–484.

Breiman, L., Friedman, J., Stone, C.J. and Olshen, R., *Classification and Regression Trees*, 1984 (Wadsworth and Brooks: Belmont).

Cho, V., Wuthrich, B. and Zhang, J., Text processing for classification. *J. Comput. Intell. Finance*, 1999, **7**, 6–22.

Chua, C., Milosavljevic, M. and Curran, J.R., A sentiment detection engine for internet stock message boards. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pp. 89–93, 2009 (Sydney, Australia).

Collins, M., Schapire, R.E. and Singer, Y., Logistic regression. AdaBoost and Bregman distances. *Machine Learn.*, 2004, **48**, 253–285.

Creamer, G. and Freund, Y., Predicting performance and quantifying corporate governance risk for Latin American ADRs and banks. In *Proceedings of the Second IASTED International Conference Financial Engineering and Applications*, pp. 91–101, 2004 (Acta Press: Cambridge, MA).

Creamer, G. and Freund, Y., Using AdaBoost for equity investment scorecards. In *Proceedings of the Machine Learning in Finance Workshop in NIPS 2005*, 2005 (Whistler, BC).

Creamer, G. and Freund, Y., A boosting approach for automated trading. *J. Trading*, 2007, **2**, 84–95.

Creamer, G. and Stolfo, S., A link mining algorithm for earnings forecast and trading. *Data Min. Knowl. Disc.*, 2009, **18**, 419–445.

Creamer, G.G., Ren, Y. and Nickerson, J.V., Impact of dynamic corporate news networks on asset return and volatility. In *2010 IEEE International Conference on Social Computing/IEEE International Conference on Privacy, Security, Risk and Trust*, Alexandria, VA, pp. 809–814, 2013.

Das, D., Schneider, N., Chen, D. and Smith, N.A., Probabilistic frame-semantic parsing. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Los Angeles, CA, 2010.

de Nooy, W., Mrvar, A. and Batagelj, V., *Exploratory Social Network Analysis with Pajek*, 2005 (Cambridge University Press: New York).

Decker, K., Sycara, K. and Zeng, D., Designing a multi-agent portfolio management system. In *Proceedings of the AAAI Workshop on Internet Information Systems*, 1996 (AAI Press: Menlo Park, CA).

Devitt, A. and Ahmad, K., Sentiment polarity identification in financial news: A cohesion based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 984–991, 2007 (Association for Computational Linguistics: Prague, Czech Republic).

Dhar, V. and Chou, D., A comparison of nonlinear methods for predicting earnings surprises and returns. *IEEE Trans. Neural Netw.*, 2001, **12**, 907–921.

Domingos, P. and Richardson, M., Mining the network value of customers. In *Proceedings of the KDD '01: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, pp. 57–66, 2001 (ACM: New York, NY, USA).

Fawcett, T. and Provost, F., Activity Monitoring: Noticing interesting changes in behavior. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99)*, San Diego, CA, USA, 15–18 August, 1999, pp. 53–62, 1999 (ACM: New York).

Fillmore, C.J., Frame semantics. In *Linguistics in the Morning Calm*, edited by T.L.S. of America, 1982 (Hanshin Publishing Co.: Seoul, South Korea).

Forman, G., An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 2003, **3**, 1289–1305.

Freeman, L., Centrality in social networks conceptual clarification. *Soc. Netw.*, 1979, **1**, 215–39.

Freund, Y. and Mason, L., The alternating decision tree learning algorithm. In *Proceedings of the Machine Learning: Proceedings Sixteenth International Conference*, pp. 124–133, 1999 (Morgan Kaufmann Publishers Inc.: San Francisco, CA).

Freund, Y. and Schapire, R.E., A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 1997, **55**, 119–139.

Friedman, J., Hastie, T. and Tibshirani, R., Additive logistic regression: A statistical view of boosting. *Ann. Stat.*, 2000, **38**, 337–374.

Goldberg, H.G., Kirkland, J.D., Lee, D., Shyr, P. and Thakker, D., The NASD securities observation, news analysis and regulation system (SONAR). In *Proceedings of the IAAI 2003*, 2003 (Acapulco, Mexico).

Hagenau, M., Liebmann, M., Hedwig, M. and Neumann, D., Automated news reading: Stock price prediction based on financial news using context-specific features. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, 2012 (Big Island).

Haider, S.A. and Mehrotra, R., Corporate news classification and valence prediction: A supervised approach. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2011)*, pp. 175–181, 2011 (Association for Computational Linguistics: Portland, Oregon).

Hill, S., Provost, F. and Volinsky, C., Network-based marketing: Identifying likely adopters via consumer networks. *Stat. Sci.*, 2006, **21**, 256–276.

Idzorek, T., A step-by-step guide to the Black–Litterman model. Working Paper, 2009.

Jegadeesh, N., Kim, J., Krische, S.D. and Lee, C.M.C., Analyzing the analysts: When do recommendations add value? *J. Finance*, 2004, **59**, 1083–1124.

Khrishnan, H. and Mains, N., The two-factor Black–Litterman model, *Risk*, 2005, 69–73.

Kirkland, J.D., Senator, T.E., Hayden, J.J., Dybala, T., Goldberg, H.G. and Shyr, P., The NASD regulation advanced detection system (ADS). *AI Mag.*, 1999, **20**, 55–67.

Lavrenko, V., Schmill, M., Lawrie, D., Oglivie, P., Jensen, D. and Allan, J., Language models for financial news recommendation. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*, McLean, VA, 6–11 November, 2000, pp. 389–396, 2000 (ACM: New York).

Leskovec, J., Adamic, L.A. and Huberman, B.A., The dynamics of viral marketing. In *Proceedings of the EC '06: Proceedings of the 7th ACM Conference on Electronic Commerce*, Ann Arbor, MI, USA, pp. 228–237, 2006 (ACM: New York, NY, USA).

Markowitz, H., Portfolio selection. *J. Finance*, 1952, **7**, 77–91.

Markowitz, H., *Portfolio Selection: Efficient Diversification of Investments*, 1959 (John Wiley & Sons: New York).

Richardson, M. and Domingos, P., Markov logic networks. *Mach. Learn.*, 2006, **62**, 107–136.

Ruiz, E.J., Hristidis, V., Castillo, C., Gionis, A. and Jaimes, A., Correlating financial time series with micro-blogging activity. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, pp. 513–522, 2012 (ACM: Portland, OR).

Schumaker, R.P. and Chen, H., Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Trans. Inform. Syst.* 2009, **27**, 12:1–12:19.

Senator, T.E., Link mining applications: Progress and challenges. *SIGKDD Explor.*, 2005, **7**, 76–83.

Seo, Y.W., Giampapa, J. and Sycara, K., Financial news analysis for intelligent portfolio management. Technical report CMU-RI-TR-04-04, Robotics Institute, Carnegie Mellon University, 2004.

Sparrow, M.K., The application of network analysis to criminal intelligence: An assessment of the prospects. *Soc. Netw.*, 1991, **13**, 251–274.

Steyvers, M. and Griffiths, T., Probabilistic topic models. In *Latent Semantic Analysis: A Road to Meaning*, edited by T. Landauer, D. McNamara, S. Dennis and W. Kintsch, 2007 (Erlbaum: Mahwah, NJ).

Thomas, J.D., News and trading rules. PhD Thesis, Carnegie Mellon University, 2003.

Walters, J., The Black–Litterman model in detail. Working Paper, 2009.

Watts, D. and Strogatz, S., Collective dynamics of small world networks. *Nature*, 1998, **393**, 440–442.

Wuthrich, B., Permunetilleke, D., Leung, S., Cho, V., Zhang, J. and Lam, W., Daily prediction of major stock indices from textual www data. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, New York, 27–31 August, 1998, pp. 364–368, 1998 (AAAI Press: New York).

Xie, B., Passonneau, R.J., Wu, L. and Creamer, G., Semantic frames to predict stock price movement. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013 (Association for Computational Linguistics: Sofia, Bulgaria).

Appendix 1. Investment signals used for prediction

I do not include firm-specific subscripts in order to clarify the presentation. Subscript q refers to the most recent quarter for which an earnings announcement was made. The fundamental variables are calculated using the information of the previous quarter (SUE, SG, TA and CAPEX). My notation is similar to the notation used by Jegadeesh *et al.* (2004).

Variable	Description	Calculation detail
SECTOR	Two-digit sector classification according to the Global Industrial Classification Standards (GICS) code	Energy 10, Materials 15, Industrials 20, Consumer Discretionary 25, Consumer Staples 30, Health Care 35, Financials 40, Information Technology 45, Telecommunication Services 50 and Utilities 55
Price momentum		
CAR1	Cumulative abnormal return for the preceding six months since the earnings announcement day	$[\Pi_{t=m-6}^{m-1}(1+R_t)-1]-[\Pi_{t=m-6}^{m-1}(1+R_{tw})-1]$, where R_t is the return in month t, R_{tw} is the value weighted market return in month t, and m is the last month of quarter
CAR2	Cumulative abnormal return for the second preceding six months since the earnings announcement day	$[\Pi_{t=m-12}^{m-7}(1+R_t)-1]-[\Pi_{t=m-6}^{m-1}(1+R_{tw})-1]$
Analysts variables		
ANFOR (ANFORLAG)	Number of analysts predicting that earnings surprise increase (lagged value)	
CONSENSUS	Mean of earnings estimate by financial analysts	
FELAG	Lagged forecast error	$\frac{\text{CONSENSUS}_q-\text{EPS}_q}{\lvert\text{CONSENSUS}_q\rvert+\lvert\text{EPS}_q\rvert}$ (Dhar and Chou 2001) where EPS is earnings per share
Earnings momentum		
FREV	Analysts earnings forecast revisions to price	$\sum_{i=0}^{5}\frac{\text{CONSENSUS}_{m-i}-\text{CONSENSUS}_{m-i-1}}{P_{m-i-1}}$ where P_{m-1} is the price at end of month $m-1$, and i refers to the previous earnings revisions
SUE	Standardized unexpected earnings	$\frac{(EPS_q-EPS_{q-4})}{\sigma_t}$ where EPS is the earnings per share, and σ_t is the standard deviation of EPS for previous seven quarters
Growth indicators		
LTG	Mean of analysts' long-term growth forecast	
SG	Sales growth	$\frac{\sum_{t=0}^{3}\text{Sales}_{q-t}}{\sum_{t=0}^{3}\text{Sales}_{q-4-t}}$
Firm size		
SIZE	Market cap (natural log)	$ln(P_q\ shares_q)$ where $shares_q$ are outstanding shares at end of quarter q
Fundamentals		
TA	Total accruals to total assets	$\frac{\triangle\text{C.As.}_q-\triangle Cash_q-(\triangle\text{C.Lb.}_q-\triangle\text{C.Lb.D}_q)-\triangle\text{T}_q-\text{D\&A}_q}{\frac{(\text{T.As.}_q-\text{T.As.}_{q-4})}{2}}$ where $\triangle X_q = X_q - X_{q-1}$ and C.As., C.Lb., C.Lb.D., T,D&A and T.As. stands for current assets, current liabilities, debt in current liabilities, deferred taxes, depreciation and amortization and total assets, respectively.
CAPEX	Rolling sum of capital expenditures to total assets	$\frac{\sum_{t=0}^{3}\text{capital expenditures}_{q-t}}{(T.As._q-T.As._{q-4})/2}$
Valuation multiples		
BP	Book to price ratio	$\frac{\text{book value of common equity}_q}{\text{market cap}_q}$, where $\text{market cap}_q = P_q\ \text{shares}_q$
EP	Earnings to price ratio (rolling sum of EPS of the previous four quarters deflated by prices)	$\frac{\sum_{t=0}^{3}EPS_{q-t}}{P_q}$
Social networks		
$deg(v_i)$	Degree centrality or degree: number of edges incidents in vertex v_i	$\sum_j a_{ij}$, where a_{ij} is an element of the adjacent matrix A
$C_c(v_i)$	Closeness centrality (normalized): inverse of the average geodesic distance from vertex v_i to all other vertices	$\frac{n-1}{\sum_j d_{ij}}$, where d_{ij} is an element of the geodesic distance matrix D (Freeman 1979, Borgatti and Everett 2006)
$B_c(v_i)$	Betweenness centrality: proportion of all geodesic distances of all other vertices that include vertex v_i	$\sum_i\sum_j\frac{g_{kij}}{g_{kj}}$, where g_{kij} is the number of geodesic paths between vertices k and j that include vertex i, and g_{kj} is the number of geodesic paths between k and j Freeman (1979)
CC_i	Clustering coefficient: cliquishness of a particular neighbourhood or the proportion of edges between vertices in the neighbourhood of v_i divided by the number of edges that could exist between them (Watts and Strogatz 1998)	$\frac{2\lvert\{e_{ij}\}\rvert}{deg(v_i)(deg(v_i)-1)} : v_j \in N_i,\ e_{ij} \in E$, where each vertex v_i has a neighbourhood N defined by its immediately connected neighbours: $N_i = \{v_j\} : e_{ij} \in E$.
CC_i'	Normalized clustering coefficient	$\frac{deg(v_i)}{MaxDeg}CC_i$, where MaxDeg is the maximum degree of vertex in a network (de Nooy *et al*. 2005)
Labels		
LABELFE	Label of forecast error (FE)	1 if $CONSENSUS \geq EPS$ (current quarter), -1 otherwise
LABELCAR	Label of cumulative abnormal return (CAR)	1 if $CAR_{m+1} \geq 0$, -1 otherwise, where CAR_{m+1} refers to the CAR of the month that follows the earnings announcement