European Journal of Operational Research 257 (2017) 243–250

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

On the risk prediction and analysis of soft information in finance reports

Ming-Feng Tsai[a], Chuan-Ju Wang[b,*]

[a] Department of Computer Science and Program in Digital Content and Technology, National Chengchi University, No. 64, Sec. 2, Zhinan Rd., Taipei 116, Taiwan

[b] Research Center for Information Technology Innovation, Academia Sinica, No. 128 Academia Road, Sec. 2, Taipei 115, Taiwan

ARTICLE INFO

Article history:
Received 29 September 2015
Accepted 29 June 2016
Available online 15 July 2016

Keywords:
Finance
Risk prediction
Text mining
Sentiment analysis

ABSTRACT

We attempt in this paper to utilize soft information in financial reports to analyze financial risk among companies. Specifically, on the basis of the text information in financial reports, which is the so-called soft information, we apply analytical techniques to study relations between texts and financial risk. Furthermore, we conduct a study on financial sentiment analysis by using a finance-specific sentiment lexicon to examine the relations between financial sentiment words and financial risk. A large collection of financial reports published annually by publicly-traded companies is employed to conduct our experiments; moreover, two analytical techniques – regression and ranking methods – are applied to conduct these analyses. The experimental results show that, based on a bag-of-words model, using only financial sentiment words results in performance comparable to using the whole texts; this confirms the importance of financial sentiment words with respect to risk prediction. In addition to this performance comparison, via the learned models, we draw attention to some strong and interesting correlations between texts and financial risk. These valuable findings yield greater insight and understanding into the usefulness of soft information in financial reports and can be applied to a broad range of financial and accounting applications.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

The great amounts of data in today's environment make it more and more important to determine how to discover useful insights for improved decision-making. These discovered insights can result in the ability to take advantage of opportunities, minimize risks, and control costs. Big data analytics refers to techniques for exploring, discovering, and making data-driven decisions in the context of abundant data. These techniques include efforts toward using new analytic methods on either new data or data that has been combined in new ways.

Due to the prevalence of big data analytics, in recent years researchers have started to focus on analyzing new types of information. In finance, there are typically two kinds of information (Petersen, 2004): soft information, which usually refers to text, including opinions, ideas, and market commentary; and hard information, that is, numbers such as financial measures and historical prices. In contrast to previous works which use only hard information in the modeling of financial risk, in this paper we aim to incorporate soft information to study financial risk among companies.

Financial risk is the chance that a chosen investment instrument (e.g., stock) will lead to a loss. In finance, volatility is a common empirical measure of risk. Our main focus in this paper is to apply sentiment analysis to the task of risk prediction in an attempt to discover useful insights. In this study, we use a finance-specific sentiment lexicon to model the relations between sentiment information and financial risk; in specific, two analytic techniques are adopted: regression and ranking methods, and the texts are the annual SEC[1]-mandated financial reports. For the regression task, we attempt to predict stock return volatility via soft textual information. However, according to Kogan, Levin, Routledge, Sagi, and Smith (2009), it is considered difficult to thus predict real-world quantities using text information only; this is probably due to the huge amount of noise within text. Therefore, we propose solving this noise problem by using ranking techniques. Specifically, we first split the volatilities of company stock returns within a given year into several relative risk levels, and then we apply

* Corresponding author.
E-mail addresses: mftsai@nccu.edu.tw (M.-F. Tsai), cjwang@citi.sinica.edu.tw (C.-J. Wang).

[1] Securities and Exchange Commission.

http://dx.doi.org/10.1016/j.ejor.2016.06.069
0377-2217/© 2016 Elsevier B.V. All rights reserved.

ranking techniques to rank the companies according to their relative risk levels. From the experimental results, we observe that, when trained on the finance-specific sentiment lexicon only, both regression and ranking models yield performance comparable to those trained on the original texts, even though the word dimension is reduced considerably, from hundreds of thousands to around only 1500. This indicates that finance-specific sentiments are the most crucial ingredients in financial reports. In addition, we also conduct analyses on the resultant models; this yields more insight and understanding into the impact of soft information in financial reports.

In addition to the proposed techniques, this paper also presents a web-based information system for financial report analysis and visualization to bridge the gap between technical results and useful interpretations.[2] With the system and our analyzed results, both academics and practitioners can more easily capture useful insights and understand the impact of soft information in financial reports. One potential application of the analyzed soft information is to help banks improve their credit-risk assessment, in particular their approach to qualitative assessment.[3] Moreover, practitioners such as fund managers can utilize the learned high-risk sentiment keywords to assist in designing their own investment strategies. For accounting research also, understanding the soft information in financial reports is a vital task, because the soft information can provide a very helpful context for understanding financial data and testing interesting economic hypotheses (Li, 2010). Therefore, it can be said that this study can be applied to a broad range of financial and accounting applications.

The remainder of this paper is organized as follows. In Section 2, we present related past work and outline our aims. We then describe in Section 3, how we accomplish our analysis: the definition of the risk measure, the mechanism of risk-level splitting, the financial sentiment lexicon, and the problem formulation. In Section 4, we present the details of our experimental settings and experimental results. In Section 5, we provide discussion and analysis, after which we conclude the paper.

2. Related work

In finance, there are typically two kinds of information: soft and hard information (Petersen, 2004). Soft information usually refers to textual information, including opinions, ideas, and market commentary, and hard information refers to numerical information such as historical time series of stock prices. Most financial studies related to risk analysis are based on hard numerical information, especially time series modeling (e.g., Armano, Marchesi, & Murru, 2005; Bodyanskiy & Popov, 2006; Christoffersen & Diebold, 2000; Chu, Santoni, & Liu, 1996; Dash, Hanumara, & Kajiji, 2003; Fu, 2011; Hung, 2009; Laih, 2014; Lee & Tong, 2011; Wu, Chen, & Olson, 2014; Yümlü, Gürgen, & Okay, 2005; Wong, Xia, & Chu, 2010). In natural language processing, some have used regression to predict continuous quantities. For instance, Mcauliffe and Blei (2007) predicted movie reviews and popularity from text via latent "topic" variables, and Lavrenko et al. (2000) used language models to analyze influences between text and time-series financial data (stock prices). In addition, in information retrieval, in recent years there have also been attempts to use learning-based methods to solve the text ranking problem (e.g., Burges et al., 2005; Freund, Iyer, Schapire, & Singer, 2003; Joachims, 2006), which has subsequently brought to the fore the topic of "learning to rank" in the fields of information retrieval and machine learning.

Some researchers have focused on mining financial reports or news (e.g., Balakrishnan, Qiu, & Srinivasan, 2010; Blasco, Corredor, Del Rio, & Santamarıa, 2005; Groth & Muntermann, 2011; Huang & Li, 2011; Kogan et al., 2009; Leidner & Schilder, 2010; Lin, Lee, Kao, & Chen, 2008; Schumaker & Chen, 2009). Lin et al. (2008) used a weighting scheme to combine both qualitative and quantitative features of financial reports, and then proposed a method to predict short-term stock price movements. They used a hierarchical agglomerative clustering (HAC) method with K-means updating to improve the purity of the prototypes of financial reports, and then used the generated prototypes to predict stock price movements. Other research has focused on predicting risk from financial reports, for instance (Leidner & Schilder, 2010), in which the text mining was used to detect risks within a company, and then classify the detected risk into several types. The above two studies both used classification to mine financial reports. In 2009, Kogan et al. (2009) applied a regression approach to predict stock return volatilities of companies via their financial reports; specifically, the support vector regression (SVR) model was applied to mine the text information. Also, two state-of-the-art studies on textual information in MD&A disclosures have been conducted by Ball, Hoberg, and Maksimovic (2015), Frankel, Jennings, and Lee (2015); the first study points out that the content of the MD&A can be systematically adopted to explain the valuation of firms, whereas the second utilizes MD&A disclosures to predict current-year firm-level accruals via SVR.

Furthermore, following the explosion of sentiment information from social web sites, blogs, and online forums, sentiment analysis has emerged as a popular research area in computational linguistics (Mohammad & Turney, 2010; Narayanan, Liu, & Choudhary, 2009). In general, sentiment analysis attempts to determine author attitudes about given topics: this could include the author's judgments or evaluations, the author's emotional state when writing, or the author's intended emotional communication to readers. The growing importance of sentiment analysis applied to finance raises many research and practical issues, such as "Why is sentiment analysis important?" In finance, there have been several studies (e.g., Garcia, 2013; Loughran & McDonald, 2011; Price, Doran, Peterson, & Bliss, 2012) that used textual analysis to examine the sentiment of numerous news items, articles, financial reports, and tweets about public companies. For most sentiment analysis algorithms, the sentiment lexicon is the most important resource and has yielded improved results and analysis (Feldman, 2013). However, past works usually used general sentiment lexicons for analysis. As mentioned in Loughran and McDonald (2011), a general purpose sentiment lexicon can be prone to misclassify common words in financial texts; as shown in their work, almost three-fourths of the words in financial reports, which are identified as negative by the widely used Harvard Psychosociological Dictionary, are typically not considered negative in financial contexts.

In this paper we aim to apply the analytical techniques of regression and ranking methods to study the relations between texts and financial risk; moreover, we also conduct a study on sentiment analysis using a finance-specific sentiment lexicon. Via the experimental results, we attempt to identify interesting correlations between texts and financial risk in order to provide insights and understanding into the impact of soft information in financial reports.

3. Methodology

3.1. Stock return volatility

In finance, *volatility* is a common risk metric defined as the standard deviation of a stock's returns over a period of time. Historical volatilities can be derived from time series of past market

[2] The system is available at http://clip.csie.org/10K/.

[3] Please refer to http://www.mckinsey.com/business-functions/risk/our-insights/ratings-revisited-textual-analysis-for-better-risk-management for more details.

prices. This paper uses the historical volatility of a company's stock prices as a proxy for financial risk.

Let S_t be the price of a stock at time t. Holding the stock for one period from time $t-1$ to time t results in a simple net return of

$$R_t = \frac{S_t}{S_{t-1}} - 1$$

Tsay (2005). Therefore, the volatility of returns for a stock from time $t-n$ to t can be defined as

$$v_{[t-n,t]} = \sqrt{\frac{\sum_{i=t-n}^{t} (R_i - \bar{R})^2}{n}}, \tag{1}$$

where $\bar{R} = \sum_{i=t-n}^{t} R_i/(n+1)$. Note that in this paper we use the daily returns of the stock prices.

3.2. Risk-level splitting mechanism

We now proceed to introduce the risk-level splitting mechanism by which we classify the volatilities of n stocks into $2\ell+1$ risk levels, where $n, \ell \in \{1, 2, 3, \ldots\}$. Let m be the sample mean and s be the sample standard deviation of the logarithms of the volatilities of n stocks (denoted as $\ln(v)$).[4] The distribution over $\ln(v)$ across companies approximates a bell shape (Kogan et al., 2009). Therefore, given a volatility v, we derive the risk level r as

$$r = \begin{cases} \ell - k & \text{if } \ln(v) \in (a, m - usk], \\ \ell & \text{if } \ln(v) \in (m - us, m + us), \\ \ell + k & \text{if } \ln(v) \in [m + usk, b), \end{cases} \tag{2}$$

where $a = m - us(k+1)$ when $k \in \{1, \ldots, \ell-1\}$, $a = -\infty$ when $k = \ell$, $b = m + su(k+1)$ when $k \in \{1, \ldots, \ell-1\}$, $b = \infty$ when $k = \ell$, and u is a positive real number. For example, with $\ell = 2$ and $u = 1$, there are 5 risk levels (i.e., 0,1,2,3,4):

$$r = \begin{cases} 0 & \text{if } \ln(v) \in (-\infty, m - 2s], \\ 1 & \text{if } \ln(v) \in (m - 2s, m - s], \\ 2 & \text{if } \ln(v) \in (m - s, m + s), \\ 3 & \text{if } \ln(v) \in [m + s, m + 2s), \\ 4 & \text{if } \ln(v) \in [m + 2s, \infty). \end{cases} \tag{3}$$

Note that r stands for the *relative* risk among n stocks; for instance, a stock with $r = 4$ is much riskier than one with $r = 0$.

3.3. Financial sentiment lexicon

For most sentiment analysis algorithms, the sentiment lexicon is the most crucial resource. As mentioned in Loughran and McDonald (2011), a general-purpose sentiment lexicon can misclassify common words in financial texts. As shown in their paper, almost three-fourths of the words in the 10-K financial reports from year 1994 to 2008 are identified as negative by the widely used Harvard Psychosociological Dictionary and yet are typically not considered negative in financial contexts.

In this paper, we use a finance-specific lexicon that consists of the six word lists provided by Loughran and McDonald (2011) to analyze the relations between these sentiment words and financial risk. The six lists are:[5]

1. Fin-Neg: negative business terminologies (e.g., *deficit*, *default*).
2. Fin-Pos: positive business terminologies (e.g., *achieve*, *profit*).
3. Fin-Unc: words denoting uncertainty, with emphasis on the general notion of imprecision rather than exclusively focusing on risk (e.g., *appear*, *doubt*).
4. Fin-Lit: words reflecting a propensity for legal contest or, per our label, litigiousness (e.g., *amend*, *forbear*).
5. MW-Strong: words expressing strong levels of confidence (e.g., *always*, *must*).
6. MW-Weak: words expressing weak levels of confidence (e.g., *could*, *might*).

3.4. Problem formulation

In the following two sections we formulate the two analytic techniques – regression and ranking – which are used to solve the financial risk prediction and analysis problem.

3.4.1. Regression task

Given a collection of financial reports $D = \{\mathbf{d_1}, \mathbf{d_2}, \ldots, \mathbf{d_n}\}$ in which each $\mathbf{d_i} \in \mathbb{R}^p$ (that is, each document is a p-dimensional vector) is associated with a company c_i, we seek to predict the company's future risk, which is characterized by its volatility v_i. Such a prediction can be defined by a parameterized function f as

$$\hat{v}_i = f(\mathbf{d_i}; \mathbf{w}). \tag{4}$$

The goal is to learn a p-dimensional vector $\mathbf{w}$ given the training data $T = \{(\mathbf{d}_i, v_i) | \mathbf{d}_i \in \mathbb{R}^p, v_i \in \mathbb{R}\}$.

Support vector regression (SVR) (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997) is a popular technique for training this type of regression model. SVR is trained by solving the following optimization problem:

$$\min_{\mathbf{w}} V(\mathbf{w}) = \frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle + \frac{C}{n} \sum_{i=1}^{n} \max\left(|v_i - f(\mathbf{d}_i; \mathbf{w})| - \epsilon, 0\right),$$

where C is a regularization constant and ϵ controls the training error. More details about SVR can be found in Schölkopf and Smola (2001).

3.4.2. Ranking task

For the ranking task, given each company's financial reports, our goal is to rank companies according to the stock return volatilities. Using the aforementioned splitting mechanism, we first split each year's stock return volatilities into different risk levels; this can be considered the relative difference of risk among the companies.

After classifying the stock return volatilities (of companies) into different risk levels, the ranking task can be defined as follows: given a collection of financial reports D, we aim to rank the companies via a ranking model $f : \mathbb{R}^p \to \mathbb{R}$ such that the rank order of the set of companies is specified by the real value that the model f takes. In specific, $f(\mathbf{d_i}) > f(\mathbf{d_j})$ is taken to mean that the model asserts that $c_i \succ c_j$, where $c_i \succ c_j$ means that c_i is ranked higher than c_j; that is, the company c_i is more risky than c_j.

We adopt ranking SVM (Joachims, 2006) for this ranking problem; the purpose of ranking SVM is to minimize the number of discordant pairs while maximizing the margin of pairs. Within a given year, if the ground truth (i.e., the relative risk generated by the proposed mechanism) asserts that company c_i is more risky than c_j, the constraint of ranking SVM is $\langle \mathbf{w}, \mathbf{d_i} \rangle > \langle \mathbf{w}, \mathbf{d_j} \rangle$, where $\mathbf{w}, \mathbf{d_i}, \mathbf{d_j} \in \mathbb{R}^p$, and $\mathbf{d_i}$ and $\mathbf{d_j}$ are two p-dimensional word vectors. Then, the text ranking problem can be expressed as the following

[4] As it is standard in finance, we take the logarithm of volatilities.

[5] The lists are all available at http://www.nd.edu/~mcdonald/Word_Lists.html.

Table 1
10-K Corpora statistics. The second column shows the number of financial reports in each year. The third column denotes the numbers of unique terms after filtering and tokenization. The near doubling in average document size during 2002–2003 is possibly due to the passage of the Sarbanes–Oxley Act of 2002 in the wake of Enron's accounting scandal (and numerous others) (Kogan et al., 2009).

Year	# of documents	# of unique terms
1996	1406	19,613
1997	2260	26,039
1998	2461	29,020
1999	2524	30,359
2000	2424	30,312
2001	2596	32,292
2002	2845	38,692
2003	3611	48,513
2004	3558	50,674
2005	3474	53,388
2006	3306	51,147

constrained optimization problem:

$$\min_{\mathbf{w}} V(\mathbf{w}, \xi) = \frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle + C \sum \xi_{i,j,k}$$
$$\text{s.t.} \begin{cases} \forall (\mathbf{d_i}, \mathbf{d_j}) \in Y_1 : \langle \mathbf{w}, \mathbf{d_i} \rangle \geq \langle \mathbf{w}, \mathbf{d_j} \rangle + 1 - \xi_{i,j,1} \\ \cdots \\ \forall (\mathbf{d_i}, \mathbf{d_j}) \in Y_n : \langle \mathbf{w}, \mathbf{d_i} \rangle \geq \langle \mathbf{w}, \mathbf{d_j} \rangle + 1 - \xi_{i,j,n} \\ \forall i \forall j \forall k : \xi_{i,j,k} \geq 0, \end{cases} \quad (5)$$

where $\mathbf{w}$ is a learned weight vector, C is the trade-off parameter, $\xi_{i,j,k}$ is a slack variable, and Y_k is a year's set of pairs of financial reports.

4. Experiments

In this section we first describe the details of our experimental settings. Then, we report the experimental results of the models trained on the finance-specific sentiments only and those on original texts for the regression and ranking tasks.

4.1. Experimental settings

4.1.1. Corpora and preprocessings

In the United States, federal securities laws require publicly traded companies to disclose information on a regular basis. Form 10-K, an annual report required by the Securities and Exchange Commission (SEC), provides a comprehensive overview of the company's business and financial conditions, and includes audited financial statements. In this paper, the 10-K Corpus (Kogan et al., 2009) is used to conduct our experiments, in which only Section 7 "management's discussion and analysis of financial conditions and results of operations" (MD&A) is used because it contains the most important forward-looking statements about the company.

In our experiments, for preprocessing, all documents and all six financial sentiment word lists were stemmed using the Porter stemmer, and some stop words were also removed. Table 1 lists the numbers of documents and unique terms in each year. Table 2 shows the statistics before and after stemming for each of the six financial word lists. Note that as some words occurred in more than one word list, the number of unique stemmed sentiment words is 1546 rather than 1664.

In addition, the twelve months before and after the report volatility for each company (denoted as $v^{-(12)}$ and $v^{+(12)}$, respectively) can be calculated by Eq. (1), where the price return series can be obtained from the Center for Research in Security Prices (CRSP) US Stocks Database. For the ranking task, in order to obtain the relative risks among companies, we categorize the companies of each year into five risk levels ($\ell = 2$ and $u = 1$) via Eq. (2). Fig. 1 illustrates the risk levels from year 1996 to 2006.

Table 2
Financial lexicon statistics. The second and third columns show the number of words before and after stemming, respectively, in each of the six financial word lists. As some words occurred in more than one word list, the number of unique stemmed sentiment words is 1546 rather than 1664.

Dictionary	# of words	# of stemmed words
Fin-Neg	2349	918
Fin-Pos	354	151
Fin-Unc	291	127
Fin-Lit	871	443
MW-Strong	19	10
MW-Weak	27	15
Total	3911	1664

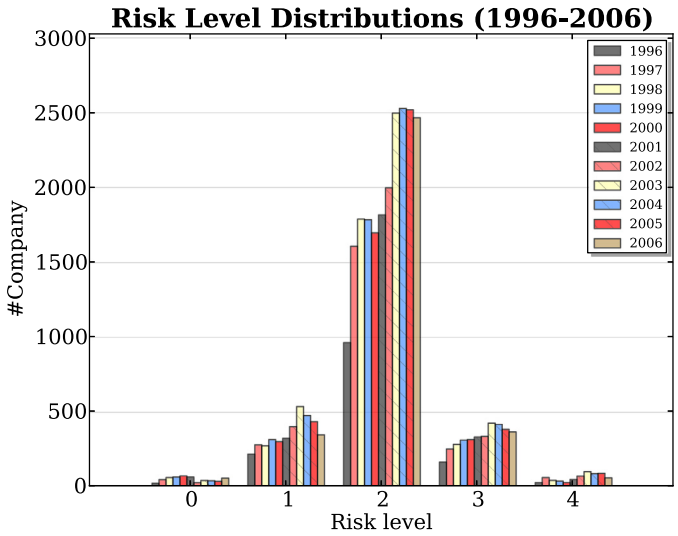

Fig. 1. Risk level distributions. The company in each year is classified into five risk levels ($l = 2$ and $u = 1$) via Eq. (2). As mentioned in Kogan et al. (2009), the distribution over $\ln(v)$ across companies approximates a bell shape.

4.1.2. Feature representation

In our experiments, for the bag-of-words model, two word features are used to represent the 10-K reports. Given a document $\mathbf{d}$, the TFIDF and LOG1P word are calculated as

- $\texttt{TFIDF}(t, \mathbf{d}) = \texttt{TF}(t, \mathbf{d}) \times \texttt{IDF}(t, \mathbf{d}) = \texttt{TC}(t, \mathbf{d})/|\mathbf{d}| \times \log(|D|/|\mathbf{d} \in D : t \in \mathbf{d}|)$,
- $\texttt{LOG1P} = \log(1 + \texttt{TC}(t, \mathbf{d}))$.

Above, $\texttt{TC}(t, \mathbf{d})$ denotes the term count of t in $\mathbf{d}$, $|\mathbf{d}|$ is the length of document $\mathbf{d}$, and D denotes the set of all documents in the corresponding year. Note that IDF is computed from the documents in a single year because the document frequency of a specific word may vary across different years. Following the work in Kogan et al. (2009), this study also uses the logarithm of the twelve-month pre-report volatility (i.e., $\log v^{-(12)}$) as an additional feature. We hereafter denote these trained models as TFIDF+ and LOG1P+.

4.1.3. Evaluation metrics

For the regression task, the performance is measured by the mean squared error (MSE) between the predicted ($\hat{v}_i^{+(12)}$) and true

Table 3
Experimental results using original texts (ALL) and only sentiment words (SEN). For the regression task, lower values are better; for the ranking task, higher values are better. Bold face denotes the best result among BL, ALL, and SEN. Notation $*$ denotes significance compared to the baseline under a permutation test ($p < 0.05$).

Task (features)		2001	2002	2003	2004	2005	2006	Average
		Mean squared error						
Regression	BL	**0.17470**	**0.16002**	0.18734	0.14421	0.13647	0.14638	**0.15086**
(LOG1P+)	ALL	0.18082	0.17175	0.17157	0.12879	0.13038	0.14287	0.15436
	SEN	0.18506	0.16367	**0.15795**	**0.12822**	**0.13029**	**0.13998**	**0.15086**
		Kendall's Tau						
	BL	0.62455	0.61973	**0.60755**	0.58616	0.59990	0.58248	0.60339
Ranking	ALL	0.62173	**0.63626**	0.58528	**0.59350**	0.59651	0.57641	0.60162
	SEN	**0.63349**	0.62280	0.60527	0.59017	**0.60273**	**0.58287**	**0.60622**$*$
(TFIDF+)		Spearman's Rho						
	BL	0.65486	0.65001	**0.63874**	0.61548	0.62857	0.60942	0.63284
	ALL	0.65271	**0.66692**	0.61662	**0.62317**	0.62531	0.60371	0.63141
	SEN	**0.66397**	0.65303	0.63646	0.61953	**0.63133**	**0.60999**	**0.63572**$*$

$(v_i^{+(12)})$ log-volatilities:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\log\left(v_i^{+(12)}\right) - \log\left(\hat{v}_i^{+(12)}\right)\right)^2,$$

where n is the number of tested companies.

For the ranking task, two rank correlation metrics are used to evaluate the performance in our experiments: Spearman's Rho (Myers, Well, & Lorch, 2010) and Kendall's Tau (Kendall, 1938). Given two ranked lists $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$,

$$\text{Rho} = 1 - \frac{6\sum (x_i - y_i)^2}{n(n^2-1)},$$

$$\text{Tau} = \frac{\#\text{concordant pairs} - \#\text{discordant pairs}}{0.5 \cdot n \cdot (n-1)}.$$

For Kendall's Tau, any pair of observations (x_i, y_i) and (x_j, y_j) is concordant if the ranks for both elements agree; that is, if both $x_i \succ x_j$ and $y_i \succ y_j$ or if both $x_j \succ x_i$ and $y_j \succ y_i$. In contrast, it is discordant if $x_i \succ x_j$ and $y_j \succ y_i$ or if $x_j \succ x_i$ and $y_i \succ y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant.

4.1.4. Parameter settings

For the regression task, we used a linear kernel with $\epsilon = 0.1$ and set the trade-off C to the default value of SVM^{light},[6] which are similar settings to those in Kogan et al. (2009). For ranking, we used a linear kernel with $C = 1$, and for all other parameters retained the default values of SVM^{Rank}.[7]

4.2. Experimental results

Table 3 tabulates the experimental results, in which the training data was composed of the financial reports in a five-year period, the year following which is the test data. For example, the reports from year 1996 to 2000 constituted the training data, and the learned model was then tested on the reports of year 2001.

Following Kogan et al. (2009), we used the logarithm of the twelve-month pre-report volatility (i.e., $\log v^{-(12)}$) as the baseline (denoted as BL hereafter). We compared the performance of the models trained on the original texts (ALL) with those trained on only sentiment words (SEN). In our experiments, the word feature LOG1P was chosen for the regression task and TFIDF for ranking, as suggested in Kogan et al. (2009) and Tsai and Wang (2013). Note that in these two studies, the models were trained on the original texts and the results are listed in the ALL row in Table 3. The bold face number in the table denotes the best result among BL, ALL, and SEN. Note that for the regression task, lower values are better, but for the ranking task, higher values are better. The notation $*$ denotes significant improvement with respect to the baseline under a permutation test ($p < 0.05$).

As shown in the table, for the two tasks, the SEN results, in most cases, are better than the ALL and BL results. Furthermore, in order to show that our ranking models with sentiment lexicon words indeed possess incremental explanatory power over the models that employ quantitative data (BL), we conducted a statistical permutation test on the resulting scores. This test confirms that the models achieved significant improvement with respect to the baseline, which corresponds to our major claim of this paper: the efficacy of the proposed ranking method and the effectiveness of the financial-specific sentiment lexicon for risk analysis with soft information. Note that although the SEN approach reduces the dimension count from hundreds of thousands to only one and half thousand, the comparable or even better results confirm that finance-specific sentiments are the most crucial ingredients in financial reports.

Our prediction results are also consistent with the findings of Kogan et al. (2009): "recency of the training set affected performance much more strongly in earlier train/test splits (2001–2003) than later ones (2004–2006)." That is, as shown in Table 3, our approaches incorporating soft information (both regression and ranking) yield better performance for reports after the passage of the Sarbanes–Oxley Act of 2002; this indicates that the more informative the report, the better prediction performance can be obtained using soft textual information.

In addition to the performance enhancement, another advantage of using soft information while predicting financial risk is that it reveals strong and interesting correlations between texts and financial risk, which is more interpretable and informative than quantitative data. These valuable findings, therefore, yield greater insight and understanding into the usefulness of soft information in financial reports and can be applied to a broad range of financial and accounting applications. Therefore, in the following Section 5, we present analysis and discussion on the top-ranked words learned by our methods.

5. Analysis

5.1. Ranking vs. regression

Fig. 2 shows the top-10 learned words from both the ranking (TFIDF+) and regression (LOGP+) models trained on sentiment

[6] http://svmlight.joachims.org/ .
[7] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html .

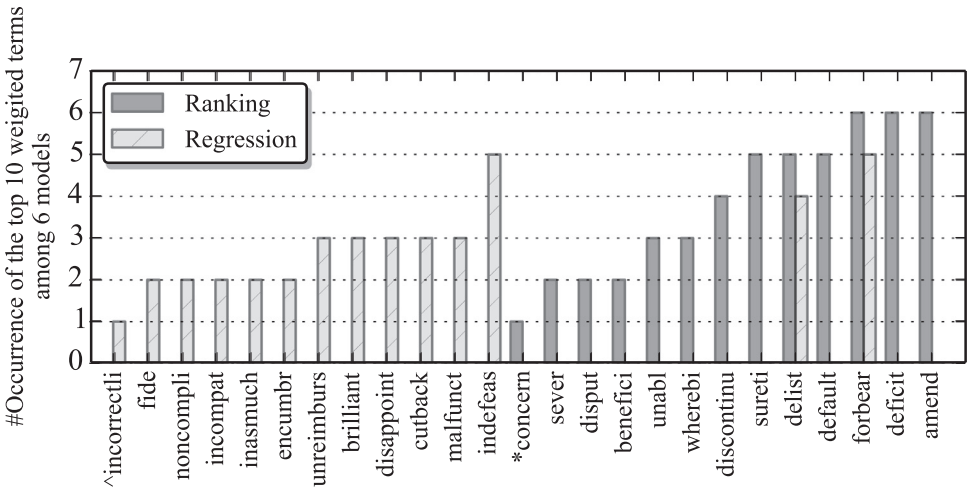

Fig. 2. Occurrence counts of the top-10 weighted terms learned via the ranking and regression tasks. Notation $*$ denotes that apart from the term *concern* there are other terms that occur only once among the six ranking models: these are *breach*, *profit*, *violat*, *regain*, *uncomplet*, *accid*, *abl*, *integr*, *doubt*, *grantor*; similarly, for the notation $\wedge$, the terms are *incorrectli*, *fault*, *nondisclosur*, *misus*, *breakag*, *defalc*, *excit*, *unclear*, *sentenc*, *overdu*, *omit*, *inforc*, *irrevoc*, *unencumb*, *further*, *variant*, *precipit*, *libel*, and *loss*.

words only (SEN); in addition, the figure also lists the accumulated numbers of these words appearing in the six corresponding regression or ranking models.

Observe that the words learned from the ranking models are much more consistent than those from the regression ones. For example, the words *amend*, *deficit*, and *forbear* appear in all of the six ranking models; in addition, there are 7 words from the ranking models that get the majority vote with more than 4 occurrences, whereas only 3 words from the regression models occur more than 4 times. On the other hand, there are 11 words from the ranking models and 20 words from the regression models that occur only once. The results shown in Fig. 2 correlate with the findings in Tsai and Wang (2013), which states that adopting ranking models to analyze the relations between financial risk and text information might be more reasonable than using regression models.

5.2. Financial sentiment terms analysis

As shown in the previous section, ranking model results are more consistent than those of regression models. Therefore, in the following discussion, we analyze words learned from the ranking models.

Table 4 tabulates the top-10 most strongly weighted terms in each of the SEN and ALL models for the ranking approach with feature TFIDF+. From these two tables, we observe that SEN-learned terms are much more consistent than those from the ALL models.

We summarize the results from Table 4 by plotting Fig. 3, in which we average the weights of each term in the six models and then rank them according to their average weights. In the figure, a single-outline circle denotes that only sentiment words were used as the training data; a double-outline circle denotes that all words in the original texts were considered during training. Color-filled circles with a term denote which sentiment word lists the term belongs to; circles with two mixed colors indicate the term belongs to two word lists. Note that the circle area is proportional to the average weight of each term.

In Fig. 3, the top 5 average-weighted words given each kind of training data are marked by numbers from 1 to 5, which correspond to the bold face numbers in Table 4. SEN training yielded the top 5 average weighted words *amend*, *deficit*, *forbear*, *delist*, and *default*, whereas those under ALL training yielded *ceg*, *nasdaq*, *gnb*, *coven*, and *forbear*; only the word *forbear* overlaps. An interesting finding is that when the models are trained on the original texts, less informative terms like *ceg* (a company name, Co-Energy Group), *nasdaq* (an American stock exchange), *gnb* (a company name, GNB Technologies) are highly ranked; however, the relation is weak between these words and financial risk. In contrast, when only sentiment words are used for training, it is more reasonable that the terms are highly related to financial risk. In addition, since the terms in the figure have been stemmed, one term may correspond to one or more words. We also list the original words from the sentiment lexicon for each top 5 average-weighted sentiment term in Fig. 3. For example, the top weighted term *amend* includes the words *amend*, *amendable*, *amendatory*, and so on.

Below we provide some original descriptions from 10-K reports that contain the top 2 weighted sentiment words in Fig. 3. Note that terms with higher weights are associated with higher financial risk. To facilitate the retrieval of the original descriptions, we further developed an information retrieval system for 10-K reports,[8] with which searches can be based on metadata or on full-text (or other content-based) indexing; the system is, therefore, of great help in extracting relevant texts and further analyzing the relationships between words and risks.

We first consider the term *amend* from the Fin-Lit list. Here is a quote from the original report:

> (from AGO, 2006 Form 10-K)
> On March 22, 2005, we *amended* the term loan agreements to, among other reasons, lower the borrowing rate by 25 basis points from LIBOR plus 2.00 percent to LIBOR plus 1.75 percent.

In finance, *amend* usually means “to change by some formal processes.” This top-ranked term reflects the fact that companies that frequently amend their policies are associated with relatively high risk.

In contrast, the term *deficit* from the Fin-Neg list means an excess of liabilities over assets, of losses over profits, or of expenditure over income in finance. Therefore, it is natural to say that a company with higher deficits might have higher risk. From the original report we have the following segment:

> (from AXS-One Inc., 2006 Form 10-K)
> At December 31, 2005, we had cash and cash equivalents of $3.6 million and a working capital *deficit* of $3.6 million which included $8.2 million of deferred revenue. The increase

[8] The system is available at http://clip.csie.org/10K/.

Table 4
Top-10 weighted terms in learned SEN and ALL models. This table lists the top-10 strongly-weighted terms in each of the SEN and ALL models for the ranking approach with the TFIDF+ feature. The number(s) in the square bracket denotes the financial list(s) that the term belongs to (zero means the term does not belong to any of the lists). Numbers in parentheses denotes the weight of the term; bold face terms denote the top-5 average weighted term in the six models and correspond to the terms in the single-outline circles numbered from 1 to 5 in Fig. 3.

Top 10 weighted terms in SEN models					
1996–2000	1997–2001	1998–2002	1999–2003	2000–2004	2001–2005
amend [4] (21.93)	**amend** [4] (21.44)	**amend** [4] (23.70)	**deficit** [2] (26.87)	**delist** [2] (25.12)	**delist** [2] (26.50)
default [2] (19.85)	**deficit** [2] (20.67)	**default** [2] (19.74)	**delist** [2] (25.13)	**amend** [4] (25.05)	**deficit** [2] (21.86)
deficit [2] (17.04)	**default** [2] (19.04)	**deficit** [2] (19.63)	**amend** [4] (23.45)	**deficit** [2] (24.24)	**forbear** [4] (20.69)
sureti [4] (15.10)	**forbear** [4] (17.23)	**forbear** [4] (17.95)	**forbear** [4] (21.59)	**forbear** [4] (20.03)	**amend** [4] (18.76)
forbear [4] (14.21)	disput [2] (16.42)	sureti [4] (16.99)	**default** [2] (19.72)	discontinu [2] (17.46)	wherebi [4] (16.05)
violat [2] (14.21)	sureti [4] (15.51)	**delist** [2] (16.90)	discontinu [2] (17.42)	wherebi [4] (16.57)	profit [1] (14.74)
disput [2] (13.81)	discontinu [2] (14.98)	concern [2] (15.97)	unabl [2] (16.07)	sureti [4] (16.37)	regain [1] (14.17)
integr [1] (13.22)	sever [2,4] (14.50)	discontinu [2] (15.81)	benefici [1,4] (15.94)	benefici [1,4] (16.14)	uncomplet [2] (14.04)
doubt [2,3] (12.99)	**delist** [2] (13.64)	breach [2,4] (14.94)	sureti [4] (15.67)	**default** [2] (16.05)	unabl [2] (13.87)
grantor [4] (12.67)	accid [2] (13.63)	sever [2,4] (14.01)	wherebi [4] (15.03)	unabl [2] (14.87)	abl [1] (13.52)
Top 10 weighted terms in ALL models					
1996–2000	1997–2001	1998–2002	1999–2003	2000–2004	2001–2005
coven [0] (13.13)	**ceg [0] (21.43)**	**ceg [0] (21.50)**	**nasdaq [0] (26.80)**	**nasdaq [0] (28.07)**	**nasdaq [0] (27.61)**
pfc [0] (12.78)	**coven** [0] (16.28)	**coven** [0] (17.80)	**ceg** [0] (20.51)	waiver [0] (21.02)	excelsior [0] (20.44)
syndic [0] (12.68)	**gnb** [0] (15.90)	**gnb** [0] (17.03)	forbear [4] (18.80)	excelsior [0] (19.43)	same [0] (19.88)
awg [0] (12.56)	ebix [0] (13.61)	waiver [0] (15.93)	shelbourn [0] (18.31)	**ceg** [0] (18.63)	**ceg** [0] (17.87)
sureti [4] (12.52)	sureti [4] (13.32)	sureti [4] (15.20)	excelsior [0] (17.17)	sureti [4] (18.50)	waiver [0] (17.34)
ebix [0] (12.41)	syndic [0] (13.28)	**default** [2] (14.87)	**gnb** [0] (16.99)	forbear [4] (16.31)	rais [0] (16.00)
amend [4] (12.14)	hearth [0] (12.97)	forbear [4] (13.67)	**coven** [0] (16.26)	**coven** [0] (16.27)	forbear [4] (15.87)
libert [0] (12.04)	stage [0] (12.89)	ebix [0] (13.53)	waiver [0] (14.93)	driver [0] (16.12)	**gnb** [0] (15.82)
stage [0] (11.90)	forbear [4] (12.81)	placement [0] (13.44)	smallcap [0] (14.67)	**gnb** [0] (15.72)	placement [0] (15.30)
special [0] (11.52)	pfc [0] (12.76)	seri [2] (13.03)	rais [0] (14.41)	rais [0] (15.41)	shelbourn [0] (14.88)

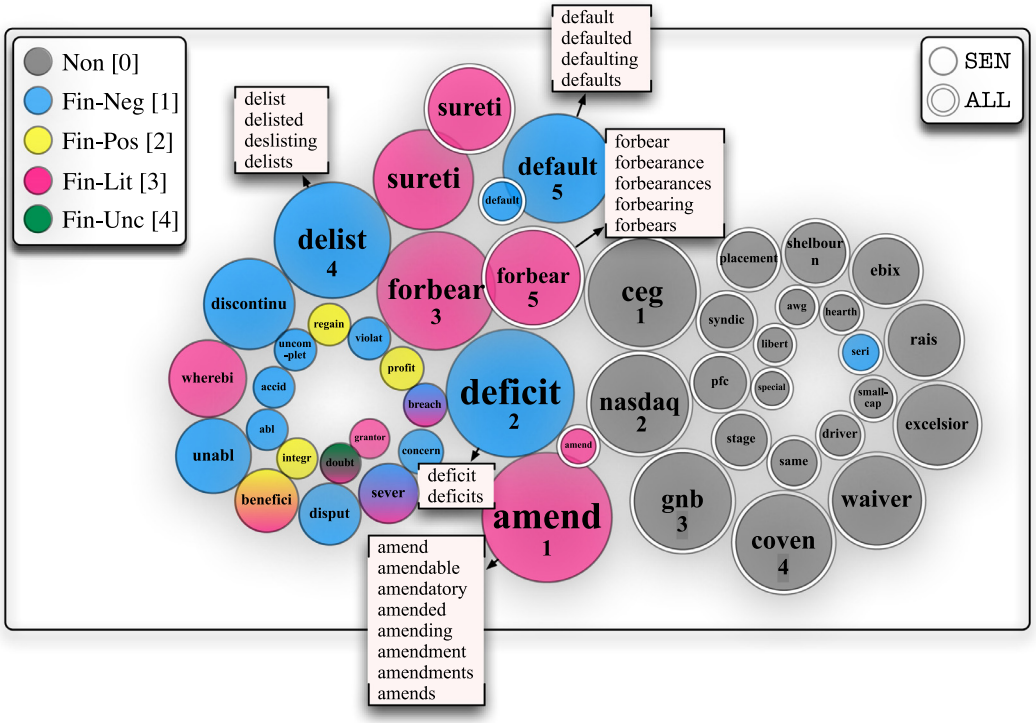

Fig. 3. Highly-weighted terms learned from the six ranking ALL and SEN models. Color-filled circles with a term denote which sentiment word lists the word belongs to; circles with two mixed colors indicate the term belongs to two word lists. Single-outline circles denote that only sentiment words from the six dictionaries (see Table 2) were used as the training data; double-outline circles denote that all words in the original texts were considered during training. The top 5 terms for the each result are marked by numbers from 1 to 5; the original words from the sentiment lexicon for each top 5 average-weighted sentiment terms are also provided. Bracketed numbers denote each type of financial list (zero means the term does not belong to any of the lists).

of the working capital *deficit* from $3.3 million at December 31, 2004 is primarily the result of a decrease in cash and decreased accounts receivable offset partially by a decrease in deferred revenue.

5.3. Summary

These analyses demonstrate that the words learned from the ranking models are much more consistent than those from the regression models. Additionally, using only sentiment words as the training data not only yields better performance than using the original texts but also provides a way to understand the relations between financial risk and financial sentiment information.

6. Conclusions

This paper identifies the importance of sentiment words in financial reports which are associated with financial risk. Using a finance-specific sentiment lexicon, we apply regression and ranking techniques to analyze the relations between sentiment words and financial risk. The experimental results show that, based on a bag-of-words model, models trained on sentiment words alone yield performance comparable to those on the original texts; this attests the importance of financial sentiment words with respect to risk prediction. In addition, the learned models also reveal strong correlations between financial sentiment words in financial reports and company risk. As a result, these findings provide more insight and understanding into the impact of financial soft textual information, especially financial sentiments, on the future risk analysis of companies. Moreover, we develop a web-based information system for financial report analysis and visualization, with which searches can be based on metadata or on full-text (or other content-based) indexing; the system is, therefore, of great help in extracting relevant texts and further analyzing the relationships between words and financial risk. This system's ease of use bridges the gap between technical results and useful interpretations, and thus renders this study more understandable to people from different fields and lends it to a broad range of financial and accounting applications.

References

Armano, G., Marchesi, M., & Murru, A. (2005). A hybrid genetic-neural architecture for stock indexes forecasting. *Information Sciences, 170*(1), 3–33.

Balakrishnan, R., Qiu, X. Y., & Srinivasan, P. (2010). On the predictive ability of narrative disclosures in annual reports. *European Journal of Operational Research, 202*(3), 789–801.

Ball, C., Hoberg, G., & Maksimovic, V. (2015). Disclosure, business change and earnings quality. Available at SSRN 2260371.

Blasco, N., Corredor, P., Del Rio, C., & Santamarıa, R. (2005). Bad news and Dow Jones make the Spanish stocks go round. *European Journal of Operational Research, 163*(1), 253–275.

Bodyanskiy, Y., & Popov, S. (2006). Neural network approach to forecasting of quasi periodic financial time series. *European Journal of Operational Research, 175*(3), 1357–1366.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., et al. (2005). Learning to rank using gradient descent. In *Proceedings of the twenty-second international conference on machine learning (ICML '05)* (pp. 89–96).

Christoffersen, P. F., & Diebold, F. X. (2000). How relevant is volatility forecasting for financial risk management? *Review of Economics and Statistics, 82*(1), 12–22.

Chu, C.-S. J., Santoni, G. J., & Liu, T. (1996). Stock market volatility and regime shifts in returns. *Information Sciences, 94*(1), 179–190.

Dash, G. H., Hanumara, C. R., & Kajiji, N. (2003). Neural network architectures for efficient modeling of FX futures options volatility. *Operational Research, 3*(1), 3–23.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems, 9*, 155–161.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM, 56*(4), 82–89.

Frankel, R. M., Jennings, J. N., & Lee, J. A. (2015). Using unstructured and qualitative disclosures to explain accruals. Available at SSRN 2563940.

Freund, Y., Iyer, R., Schapire, R. E., & Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *The Journal of Machine Learning Research, 4*, 933–969.

Fu, T.-C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence, 24*(1), 164–181.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance, 68*(3), 1267–1300.

Groth, S. S., & Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems, 50*(4), 680–691.

Huang, K.-W., & Li, Z. (2011). A multilabel text classification algorithm for labeling risk factors in sec form 10-K. *ACM Transactions on Management Information Systems, 2*(3), 18.

Hung, J.-C. (2009). A fuzzy asymmetric Garch model applied to stock markets. *Information Sciences, 179*(22), 3930–3943.

Joachims, T. (2006). Training linear SVMS in linear time. In *Proceedings of the twelfth ACM SIGKDD international conference on knowledge discovery and data mining (KDD '06)* (pp. 217–226).

Kendall, M. (1938). A new measure of rank correlation. *Biometrika, 30*, 81–93.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., & Smith, N. A. (2009). Predicting risk from financial reports with regression. In *Proceedings of the human language technologies: The 2009 annual conference of the north American chapter of the association for computational linguistics (NAACL '09)* (pp. 272–280).

Laih, Y.-W. (2014). Measuring rank correlation coefficients between financial time series: A GARCH-copula based sequence alignment algorithm. *European Journal of Operational Research, 232*(2), 375–382.

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., & Allan, J. (2000). Language models for financial news recommendation. In *Proceedings of the ninth international conference on information and knowledge management (CIKM '00)* (pp. 389–396).

Lee, Y.-S., & Tong, L.-I. (2011). Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming. *Knowledge-Based Systems, 24*(1), 66–72.

Leidner, J. L., & Schilder, F. (2010). Hunting for the black swan: risk mining from text. In *Proceedings of the ACL 2010 system demonstrations (ACLDemos '10)* (pp. 54–59).

Li, F. (2010). Textual analysis of corporate disclosures: A survey of the literature. *Journal of Accounting Literature, 29*, 143.

Lin, M.-C., Lee, A. J. T., Kao, R.-T., & Chen, K.-T. (2008). Stock price movement prediction using representative prototypes of financial reports. *ACM Transactions on Management Information Systems, 2*(3), 19:1–19:18.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance, 66*(1), 35–65.

Mcauliffe, J. D., & Blei, D. M. (2007). Supervised topic models. *Advances in Neural Information Processing Systems, 7*, 121–128.

Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 26–34).

Myers, J. L., Well, A., & Lorch, R. F. (2010). *Research design and statistical analysis*. Routledge.

Narayanan, R., Liu, B., & Choudhary, A. (2009). Sentiment analysis of conditional sentences. In *Proceedings of the 2009 conference on empirical methods in natural language processing (EMNLP '09)* (pp. 180–189). Association for Computational Linguistics.

Petersen, M. A. (2004). Information: Hard and soft. *Technical report*. Northwestern University.

Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance, 36*(4), 992–1011.

Schölkopf, B., & Smola, A. (2001). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems, 27*(2), 12:1–12:19.

Tsai, M.-F., & Wang, C.-J. (2013). Risk ranking from financial reports. In *Advances in information retrieval* (pp. 804–807). Springer.

Tsay, R. (2005). *Analysis of financial time series*. Wiley.

Wong, W. K., Xia, M., & Chu, W. (2010). Adaptive neural network model for time-series forecasting. *European Journal of Operational Research, 207*(2), 807–816.

Wu, D. D., Chen, S.-H., & Olson, D. L. (2014). Business intelligence in risk management: Some recent progresses. *Information Sciences, 256*, 1–7.

Yümlü, S., Gürgen, F. S., & Okay, N. (2005). A comparison of global, recurrent and smoothed-piecewise neural models for istanbul stock exchange (ISE) prediction. *Pattern Recognition Letters, 26*(13), 2093–2103.