Pacific-Basin Finance Journal 24 (2013) 1–21

Contents lists available at SciVerse ScienceDirect

Pacific-Basin Finance Journal

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

Revisiting early warning signals of corporate credit default using linguistic analysis ☆

Yang-Cheng Lu [a], Chung-Hua Shen [b], Yu-Chen Wei [c,*]

[a] Department of Finance, Ming Chuan University, Taipei, Taiwan
[b] Department of Finance, National Taiwan University, Taipei, Taiwan
[c] Department of Money and Banking, National Kaohsiung First University of Science and Technology, Kaohsiung, Taiwan

ARTICLE INFO

Article history:
Received 19 December 2011
Accepted 7 February 2013
Available online 16 February 2013

JEL classification:
G33
C10
G14

Keywords:
Credit default
Financial distress
Early warning
Linguistic analysis
Media
Logistic regression

ABSTRACT

We apply computational linguistic text mining (TM) analysis to extract and quantify relevant Chinese financial news in an attempt to further develop the classical early warning models of financial distress. Extending the work of Demers and Vega (2011), we propose a measure of the degree of credit default, referred to in this study as the 'distress intensity of default-corpus' (*DIDC*), and investigate the predictive power of this measure on default probability by incorporating it into the signaling model, along with the classical financial performance variables (the liquidity, debt, activity and profitability ratios). We also apply the 'naïve probability of the Merton distance to default' model (Bharath and Shumway, 2008) for our robustness analysis. A logistic regression (LR) model is constructed to better integrate the *DIDC* and financial performance variables into a more effective early warning signal model, with the incorporation of *DIDC* into the LR model revealing a significant reduction in Type I errors and an apparent increase in classification accuracy. This provides proof of the effectiveness of the additional information from TM on the financial corpus, while also confirming the predictive power of TM on credit default. The major contribution of this study stems from our potential refinement of early warning models of financial distress through the incorporation of information provided by related media reports.

☆ This work is supported by a project of the Department of Industrial Technology at the Ministry of Economic Affairs in Taiwan (Grant numbers: 99-EC-17-A-34-S1-149, 100-EC-17-A-34-S1-149 and 101-EC-17-A-34-S1-149). The support provided by the Chinese Knowledge and Information Processing Group of the Institute of Information Science and the Institute of Linguistics of Academia Sinica is also greatly appreciated.

* Corresponding author at: National Kaohsiung First University of Science and Technology, No 2 Jhuoyue Road, Nanzih District, Kaohsiung City 811, Taiwan. Tel.: +886 7 6011000x3131.
E-mail address: claireycwei@gmail.com (Y.-C. Wei).

0927-538X/$ – see front matter
http://dx.doi.org/10.1016/j.pacfin.2013.02.002

1. Introduction

Studies examining the early warning predictive signals of business default have received widespread and growing attention over recent decades, since such predictive signals have come to be regarded as extremely important for both the financial and non-financial sectors of any economy. Clearly, it is of considerable importance for firms within the non-financial sectors to be aware of the probability of default amongst their competitors; however, within the financial sector, referring particularly to the banking industry, an awareness of the probability of default will not only help to ensure appropriate capital allocation, but it can also help to reduce occurrences of non-performing loans.

It also seems obvious that investors can gain from studies on predictive signals of business default, since an awareness of such signals can clearly assist them to avoid the pursuit of poor investment targets or unwise investment in questionable assets. Signals of corporate default are also undoubtedly of value to the relevant authorities, since they can assist the various bodies to monitor the relevant industries so as to avoid any potential systematic risk.

Since the pioneering work of Altman (1968) involved only five financial ratios, the classification accuracy of bankruptcy within his models ranged from 95% in the one-year period prior to bankruptcy, to less than 50% for the three annual reporting periods prior to the default year. However, since then, the model has been extended to include, amongst other factors, industrial characteristics (Platt and Platt, 1991, 2006), the business cycle (Koopman and Lucas, 2005; Hol, 2007) and corporate governance (Johnson et al., 2000; Lee and Yeh, 2004) with the aim of further enhancing the predictive ability of the original Altman model, essentially because the model does not work so well under modern conditions.

Several alternative methodologies have also been proposed over recent decades, including logit and probit models (Ohlson, 1980; Westgaard and Wijst, 2001), multi-group hierarchical models (Doumpos et al., 2002; Doumpos and Zopounidis, 2002) and neural networks (Piramuthu, 1999; Atiya, 2001), with Dimitras et al. (1996) and Lennox (1999) comparing the performance between alternative models.It therefore seems quite clear that early warning signals of default probability have become a significant element of research, particularly in the field of finance.

Recent studies on earnings announcements additionally noting that news released in the related media provides useful information which can help to clarify pre- and post-announcement 'drift' (Vega, 2006; Demers and Vega, 2011). Within the conventional financial theories, only private or internal information is regarded as being useful in predicting abnormal returns; however, Tetlock (2007) and Tetlock et al. (2008) challenged the 'efficient markets' hypothesis by demonstrating that public news also provides valuable information of relevance to earnings announcements. We therefore argue that finding ways of exploiting the information content of media reports is worthy of further investigation.

Text mining (TM) was defined by Manning and Schütze (2002) as an effective process for the organization of unstructured textual information and the extraction of meaningful numerical indices from relevant texts. They proposed that the terms which occurred in certain documents could be extremely informative of the content of the text, and regarded the process as being particularly useful since it could render the information contained within the text accessible to various analyses.

In the present study, we extend the concept of Manning and Schütze (2002) to propose the term 'distress intensity of default-corpus' (*DIDC*) as a measure for predicting the signaling intensity on the future credit default probability of a firm, based upon the context of the 'distressed' or 'non-distressed' corpus. We investigate whether text mining, when used in conjunction with financial ratios, can actually provide additional information capable of improving default probability forecasting. Our research design also considers the market-based variable (Bharath and Shumway, 2008), referred to as the 'naïve probability of the Merton distance to default' (*NMDD*), in the final analysis of the robustness of our results.

Prior to a firm going into default, it is often found that rumors of its potential distress are already circulating within the market, resulting from issues such as 'bounced checks' or news of the resignation of the firm's treasurer. An example of this is provided by a major Taiwanese chipmaker, Procomp Informatics Ltd., since rumors were already rife throughout the market that its revolving loans were rejected by the

banks prior to the announcement of its bankruptcy.[1] Thus, if we can extract useful information from reports provided in the media, we may succeed in enhancing the conventional early warning signal models.

The characteristics of the leading indicators possessed by the news-corpus variables prior to occurrences of financial distress are invariably regarded as irrelevant noise. Thus, despite providing relevant information on the predictability of credit default, these news-corpus variables have yet to be included in any of the early warning signal models of the related studies, and this is particularly so for news in a Chinese context. We aim to fill this research gap in the present study using data on the Taiwan economy to illustrate the ways in which TM has influences in traditional Chinese society that may be of relevance to the predictive models of credit default. We begin by extending the conventional Altman model and adapting it to the special environment of Taiwan, and then apply one-, two-, three- and four-quarter-ahead prediction models, with and without consideration of the media effect, within a Chinese context.

The remainder of this paper is organized as follows. A description of our study sample and the empirical data sources is provided in Section 2, along with details of our refined measure based upon the 'distressed' and 'non-distressed' corpus obtained through computational linguistic text mining (TM) analysis. This is followed in Section 3 by a description of the methodology used to determine whether the corpus-based measure of credit default improves the predictive power of the early warning models, with a logistic regression model being constructed to confirm the predictability of our TM-based variable. The empirical results, which are presented and analyzed in Section 4, demonstrate the significant signaling power of the TM-based variable on the prediction of credit default. Finally, the conclusions drawn from this study are presented in Section 5, along with a discussion of the potential contributions and further applications of our study.

2. Data description and variable construction

Our analysis uses data on financial ratios and a quantitative indicator of media text news. We begin by describing the financial variables and definitions of the 'distressed' and 'non-distressed' samples, and then construct the 'distress intensity of default-corpus' (*DIDC*) variable based upon the application of computational linguistic text mining analysis.

2.1. Data

Our study sample includes all listed firms on the Taiwan Stock Exchange (TWSE) and GreTai Securities Market (GTSM) which experienced financial distress between the first quarter of 2001 and the fourth quarter of 2009. Any listed firms located in the financial sector are excluded from the sample since their financial structure differs from that of other firms. Where firms had more than one episode of financial distress during the period, the first episode was taken for inclusion in our sample data.

Following the sampling method proposed by Beaver (1966), matching cohorts of robust firms were collected from the full sample of listed firms capable of meeting the following criteria: (i) firms must be in the same industry as the distressed firm; and (ii) the total assets of the firm must be as close as possible to those of the distressed firm in the one-year period prior to the occurrence of financial distress, with the difference in assets being within 50%. Furthermore, in order to reduce the potential error caused by the over-sampling of distressed firms (Ohlson, 1980; Zmijewski, 1984; Platt and Platt, 2002), as opposed to using a primary scale of 1:1 or 1:2 (distressed firms divided by robust firms), we increased the number of matching samples to 1:4.

The Taiwan Stock Market is regulated by the Securities and Futures Bureau (SFB) within the Financial Supervisory Commission of the Executive Yuan. The SFB promulgates policy directives for the systematic administration and supervision of securities issuance, securities trading and futures trading, in a similar way to the role undertaken by the US Securities and Exchange Commission (SEC). All firms listed in the

[1] Procomp Informatics Ltd. defaulted on a bond payment, and despite the fact that a huge cash balance was recorded in its books, it subsequently structured for bankruptcy in June 2004. It was discovered that the executives of the firm and its overseas sales agents had colluded in overstating the firm's sales revenue, manipulating its stock price, illegally leveraging the firm's assets and arranging bonds through 'paper' companies.

TWSE and GTSM are required to routinely announce their quarterly financial statements, with the financial data subsequently being made publicly available through the Market Observation Post System (MOPS).[2]

In addition to the availability of individual financial data on the sample firms from the MOPS, there are also other representative databases in Taiwan which systematically gather financial reporting statistics. Details on the sample of distressed firms were obtained from the Taiwan Economic Journal (TEJ) database, where they were recorded under 'delisted', 'managed' or '100% margin' stocks, in conjunction with the major news items contained in the TWSE MOPS. The primary sources of the financial data used in this study were the quarterly financial statements of the listed firms within the TEJ database and the Intelligence Winner 2000 Database of InfoTimes.[3]

News reports relating to both the distressed and matching samples were collected from the InfoTimes database, a resource published by the China Times Group (a leading representative media outlet in Taiwan) which provides details of news reported in the Commercial Times and the Chinese Times.[4]

The news items that were collected for the construction of the *DIDC* measure in the present study cannot contain all media coverage, essentially because of the limitations put in place by the authorities on any unauthorized dissemination of the reports. The China Times Group is a widely-used media organization in Taiwan, since virtually all securities firms, institutions and investors either subscribe to the organization's news reports or review them online by subscribing to the China Times Group website.

In our efforts to construct a more comprehensive model for the prediction of corporate distress, we incorporated the different categories of financial variables – *Liquidity Ratio*, *Debt Ratio*, *Activity Ratio* and *Profitability Ratio* – along with a proxy for the default probability of a firm extracted from publicly available news, using data on the one-, two-, three- and four-quarter periods prior to the occurrence of financial distress.

Based upon their 'naïve probability' variable, which was constructed to approximate the functional form of the Merton (1974) 'distance to default' (*DD*) model, Bharath and Shumway (2008) found that their 'naïve probability of the Merton distance to default' (*NMDD*) model performed slightly better than the Merton *DD* model,[5] and indeed, the construction of the *NMDD* model avoids the need for the solution of any equations or the estimation of any difficult quantities. Referring to Bharath and Shumway (2008), we also incorporate the *NMDD* model in order to determine whether text mining can improve default prediction when information is incorporated conveying the stock prices of the firms.

The anticipated effects of the explanatory variables, *Liquidity Ratio*, *Debt Ratio*, *Activity Ratio*, *Profitability Ratio* and '*naïve probability of Merton distance to default*' (*NMDD*) are provided in Table 1, along with the '*distress intensity of default-corpus*' (*DIDC*) variable. We suggest that the higher the *Interest Ratio*, *Net Debt*, *Debt Ratio*, *Equity to Net Worth*, *NMDD* and *DIDC* variables, the greater the probability of credit default. The relationship between *Profitability Ratio* and the probability of default should be negative, as it seems quite clear that profitable firms are unlikely to be in a distressed situation.

2.2. Extraction of distressed and non-distressed terms using linguistic analysis

The use of intuitive 'term frequency' (*tf*) is a popular method for computing the frequency of any special terms contained within a specific document. In this study, we select those terms from the 'distressed' and 'non-distressed' classifications exhibiting high frequency as the representative characteristics of each group. tf_{cj} refers to the term frequency of the *j*th special term in the *c*th classification, where *c* is equal to 1 (2) for a 'distressed' ('non-distressed') classification.

[2] Refer to the Market Observation Post System (MOPS) website: (http://emops.twse.com.tw/ emops_all.htm) for a comprehensive explanation of the system.

[3] Interested readers can refer to the Taiwan Economic Journal (TEJ) website (http://www.finasia.biz/ensite/) for full details. InfoTimes is a subsidiary company of the China Times Group; refer to the InfoTimes website (http://www.infotimes.com.tw/new2/index.htm) for further details.

[4] The other group publishing related financial news on listed companies in the Taiwan Stock Market is the United Daily News Group. The related news reports are not incorporated in this study since the authorities place restrictions on the use of these reports; however, we suggest that by referring to the methodology adopted in the present study, financial regulators could systematically gather public media coverage to construct our proposed 'distress intensity of default-corpus' measure.

[5] Full details of the construction of the NMDD model are provided in Bharath and Shumway (2008), pages 9–10, Section 2.3 ('A naïve alternative').

Table 1
Anticipated effects of the variables on the probability of financial distress.

Variables	Definitions	Expected sign
Panel A: Liquidity Ratio		
Current Ratio	Ratio of current assets to current liabilities indicating whether a firm may be unable to meet its immediate debts	−
Quick Ratio	Ratio measuring the liquidity of a firm, calculated by taking its current assets, minus stocks, divided by current liabilities	−
Cash Flow Ratio	Ratio showing the level of cash in a firm, relative to other assets, and the use of cash in the activities of the firm	−
Receivable Turnover Ratio	Accounting measure used to quantify the effectiveness of a firm in extending credit as well as collecting debts	−
Inventory Turnover Ratio	Ratio showing how often the inventory of a firm is sold and replaced over a specific period	−
Interest Ratio	Calculated by dividing interest payments by operating revenue	+
Panel B: Debt Ratio		
Net Debt	The overall debts of a firm, calculated by netting the value of its liabilities and debts with its cash and other similar liquid assets	+
Debt Ratio	Defined as total debt divided by total assets	+
Times Interest Earned Ratio	The extent to which income is able to meet interest payments, calculated by net income + interest / interest	−
Panel C: Activity Ratio		
Total Asset Turnover Ratio	The use that a firm makes of all its assets, calculated by dividing sales by total assets	−
EPS	Earnings per share, shown as a percentage of the market price of a single share	−
Equity to Net Worth	Long-term equity investment to net worth, calculated by dividing the former by the latter	+
Net worth Turnover Ratio	The ratio of a firm's net worth turnover calculated by dividing its operating revenue by its net worth	−
EBITDA	Earnings before interest, tax, depreciation and amortization	−
Panel D: Profitability Ratio		
Profit Growth	The income of a firm after deducting expenses, indicating the percentage profit growth from one period to another	−
Profit Margin	The ratio of the profitability of a firm calculated as net income divided by revenue, or net profits divided by sales	−
Operating Profit Margin	Operating profit over a certain period divided by revenue for the same period	−
EBT/Sales	Earnings before tax divided by sales	−
ROA	Return on assets, measuring how effectively the assets of the firm are being used to generate profits	−
ROE	Return on equity, measuring profit earned for each dollar invested in the firm's stock (the bottom line measure for shareholders)	−
Panel E: Market-based variable		
NMDD	The 'naïve probability of Merton distance to default'	+
Panel F: News-corpus variable		
DIDC	The 'distress intensity of default-corpus'	+

Note: A positive sign (+) denotes a positive effect of the variable on the probability of financial distress; a negative sign (−) denotes a negative effect of the variable on the probability of financial distress.

We apply the methodology adopted in both Frakes and Baeza-Yates (1992) and Yang and Pedersen (1997), using the Chi-squared (χ^2) test to distinguish between the representative special terms, based upon the following formula:

$$\chi^2(c,j) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \tag{1}$$

where A (B) refers to the jth (non-jth) term frequencies shown in the cth classification; C (D) refers to the jth term (non-jth) term frequencies shown in the non-cth classification; and N denotes the all-term frequencies in the documents.

The Chi-squared (χ^2) test, which evaluates whether c and j are independent or correlated, can be compared to a Chi-square distribution with one degree of freedom to judge the level of extremeness. If the Chi-squared test is higher than the critical value, thereby implying the rejection of independence, then the selected jth special term can be classified as the representative term in the cth classification.

2.3. Relative weighting of the distressed and non-distressed terms

This study adopts 'entropy' theory to explain that the special terms within the classification could transmit different levels of decision information; this approach has been widely used in computational linguistic analysis as a measure of the relative weights amongst all of the special terms. We also compare the entropy values for the different terms to determine the relative weights of these values. The entropy formula is as follows:

$$ej = -k\sum_{i=1}^{m} p_{ij} \ln p_{ij}, k = 1/\ln m; i = 1,2,3,\ldots m; j = 1,2,3,\ldots,n. \tag{2}$$

where m is the number of firms; p_{ij} is the jth term frequency shown in the ith firm; and e_j is the entropy value of the jth term.

We then go on to calculate the values of the relative weights amongst all of the special terms.

$$w_j = \left(1-e_j\right)/n-\sum_{j=1}^{n} e_j, j = 1,2,3,\ldots,n. \tag{3}$$

where n is the number of special terms. Thus, we can calculate a value for the relative weights in both the financially distressed and non-distressed classifications.

The key terms for distressed and non-distressed scenarios may be invariant over time; for example, the characteristic terms for distressed firms include 'downgrade', 'default', 'refund', and so on, and these terms may appear in the related news regardless of whether they occur in earlier or later years. The difficulty involved in the collection of key terms lies in the fact that these terms should be general principles, and therefore the collection of these terms needs to be based upon a huge volume of financial reports, which is why the identification of the special terms for the distressed and non-distressed groups is based upon the entire sample period.

The results are presented in Table 2, from which we can see that when appearing in both the distressed and non-distressed groups, some of the special terms are found to have different weights; this is essentially because those special terms with greater weights are generally viewed as negative (positive) signals in the distressed (non-distressed) groups.

Following the collation of the special terms and weights respectively based upon the Chi-squared test and the Entropy method, the lists of terms were checked and the formal terms retained, since the typical terms for the 'distressed' and 'non-distressed' groups should have general characteristics. The total number of terms was 110 for the distressed group and 129 for the non-distressed group.[6] We propose that the key terms for these groups could be applied to future studies.

2.4. Quantifying the degree of credit default

If we are to determine an efficient default probability profile, then the special terms extracted through the use of computational linguistic analysis for both the distressed and non-distressed groups need to be integrated in order to quantify the degree of credit default. In this study, we use a measure aimed at reflecting the degree of credit default, which we refer to as the 'distress intensity of default-corpus' (*DIDC*).

[6] Due to space limitations, the details of special terms presented in Table 2 include only those terms with greater weights.

Table 2
Special terms in the distressed and non-distressed groups.

Quarterly periods prior to financial distress												
0–3 months			4–6 months			7–9 months			10–12 months			
Special terms		Weight	Special terms		Weight	Special terms		Weight	Special terms		Weight	
Panel A: Distressed group												
Downgrade	(調降)	0.0035	Downgrade	(調降)	0.0050	Ex-dividend	(除權)	0.0056	Downgrade	(調降)	0.0044	
Loss	(虧損)	0.0032	Loss	(虧損)	0.0035	Downgrade	(調降)	0.0051	Ex-dividend	(除權)	0.0040	
Discipline	(處分)	0.0031	Hypothecation	(質押)	0.0024	Loss	(虧損)	0.0043	Capital formation	(增資)	0.0039	
Margin buying	(融資)	0.0026	Sales	(出貨)	0.0023	Capital formation	(增資)	0.0039	Loss	(虧損)	0.0038	
Order	(訂單)	0.0025	Bounced check	(跳票)	0.0023	Accomplishment	(業績)	0.0032	Directors	(董監事)	0.0035	
Downward price limit	(跌停)	0.0023	Plunge	(重挫)	0.0022	Loss	(損失)	0.0028	Discipline	(處分)	0.0030	
Bounced check	(跳票)	0.0019	Over-fall	(跌破)	0.0019	Transformation	(轉型)	0.0017	Not enough	(不足)	0.0029	
Plunge	(重挫)	0.0014	Not good enough	(不佳)	0.0016	Enlargement	(擴大)	0.0015	Loss	(損失)	0.0028	
Capital-reducing	(減資)	0.0010	Default	(違約)	0.0013	Cum-rights	(填權)	0.0011	Demand	(需求)	0.0022	
Default	(違約)	0.0006	Refund	(退票)	0.0008	M&A	(購併)	0.0010	Capital financing	(現增)	0.0021	
Panel B: Non-distressed group												
Earnings	(營收)	0.0148	Earnings	(營收)	0.0146	Earnings	(營收)	0.0196	Earnings	(營收)	0.0183	
Growth	(成長)	0.0101	Growth	(成長)	0.0105	Growth	(成長)	0.0137	Growth	(成長)	0.0138	
Profit	(獲利)	0.0094	Sales	(出貨)	0.0052	Profit	(獲利)	0.0129	Profit	(獲利)	0.0141	
Order	(訂單)	0.0053	New height	(新高)	0.0048	Accomplishment	(業績)	0.0070	Earnings	(盈餘)	0.0067	
Demand	(需求)	0.0037	Upper price limit	(漲停)	0.0032	New high	(新高)	0.0067	Order	(訂單)	0.0065	
Rise	(上漲)	0.0025	Quantification	(量產)	0.0020	Opportunity	(機會)	0.0040	Demand	(需求)	0.0044	
Downgrade	(調降)	0.0019	Innovation	(創新)	0.0017	Increase	(提升)	0.0030	Customer	(客戶)	0.0043	
Enhancement	(挹注)	0.0016	Performance	(效益)	0.0016	Ex-dividend	(除權)	0.0020	Quantification	(量產)	0.0027	
Enlargement	(擴充)	0.0012	Shortage	(缺貨)	0.0014	Acceptance	(接獲)	0.0018	R&D	(研發)	0.0022	
Fluency	(暢旺)	0.0007	Hypothecation	(質押)	0.0003	Enhancement	(加碼)	0.0009	Downgrade	(調降)	0.0020	

Note: 'Special terms' are the terms describing the characteristics of financially distressed and non-distressed groups, with the Chinese equivalent being provided in parentheses. Due to limitations of space, the special terms presented here include only those with higher weights. A complete list of the terms for each group is available from the authors upon request.

In order to ensure that the *DIDC* measure can successfully identify the default probability of a firm, it can be measured by the proportion of integrated relative influences from the distressed terms over the integrated relative influences from the non-distressed terms, for firm *i*, as follows:

$$DIDC_{itk} = \frac{\sum_{r} tf^{D}_{itkr} w^{D}_{r}}{\sum_{s} tf^{ND}_{itks} w^{ND}_{s}} \tag{4}$$

$$DIDC_{iq} = mean\left(\sum_{t}\sum_{k} DIDC_{itk}\right) \tag{5}$$

where $DIDC_{itk}$ refers to the 'distress intensity of default-corpus' for the *i*th firm on date *t* in the *k*th news; $DIDC_{iq}$ denotes the average 'distress intensity of default-corpus' for the *i*th firm in quarter *q* (since the frequency used in this study is quarterly periods); w^{D}_{r} are the weights of the *r*th special 'distressed' term; w^{ND}_{s} are the weights of the *s*th special 'non-distressed' terms; tf^{D}_{itkr} is the term frequency for the *i*th firm on date *t* of the *r*th special 'distressed' term in the *k*th news; and tf^{ND}_{itks} represents the term frequency for the *i*th firm on date *t* of the *s*th special 'non-distressed' term in the *k*th news.

A higher *DIDC* indicates a relatively higher intensity of default probability, and vice versa, while the level of *DIDC* also indicates the balance of coverage in the financial media for positive versus negative news. Those news outlets that are more likely to report bad news will induce a level of *DIDC* greater than 1, even amongst non-distressed firms.

The concept of 'financial distress' originates from the study of Beaver (1966), where it was defined as "incurring a huge overdraft, default on payment of preferred stock dividends and corporate bonds, and filing for bankruptcy". It is, however, quite clear that financial distress is not limited to the inability to pay debts or simply experiencing negative net worth, business closures (Altman, 1968; Ohlson, 1980; Zmijewski, 1984), bankruptcy (Deakin, 1972; Hopwood et al., 1994), reorganization or delisting, since it also includes embezzlement, serious losses, check bouncing, credit tightening by banks and temporary suspension of all trading in the stocks of a firm, all of which provide strong signals of imminent failure.

In their development of early warning models, the majority of the prior studies have tended to use financial statements (Beaver, 1966; Altman, 1968) or other relevant information to explain the probability of financial distress; however, financial statements provide only ex post information on corporate operations, and such financial information is already disclosed. Thus, when attempting to construct an early warning financial distress model, in addition to financial reports, there is a need to consider other external information. The *DIDC* proposed in this study serves to provide external information obtained from financial news reports to improve the reliability of the current early warning models.

In the aftermath of the Asian financial crisis, numerous studies very quickly began pointing to poor 'corporate governance' as a key factor associated with financial distress through issues such as ownership concentration (Rajan and Zingales, 1998; Johnson et al., 2000) cross holdings (La Porta et al., 1999; Claessens et al., 2000; Lu and Chang, 2009), inferior management, and family or shareholder control (La Porta et al., 1999; Yeh et al., 2001; Chen and Hu, 2001; Lu and Chang, 2009).

It should, however, be noted that large amounts of information are often accumulated through news reports; thus, investors can make good use of this information by engaging in linguistic analysis, going through various documents in order to uncover the hidden characteristics of significant amounts of news items (Fayyad and Uthurusamy, 1996; Fayyad et al., 1996a, 1996b).

In the present study, we attempt to extend the traditional warning models of financial distress, probing the distress signals through the incorporation of copious amounts of financial news provided within the local media. We suggest that our proposed *DIDC* is capable of fulfilling this important role.

3. Methodology

3.1. Logistic regression model

The binary logistic regression model, initially developed by Berkson (1944), was first applied to the prediction of corporate financial distress by Ohlson (1980). The objectives in using such a logistic

Table 3
Alternative and competitive models.

Alternative models (Model (i))	Competitive models (Model (i^*))
Model (1): Liquidity ratios and *DIDC*	Model (1)*: Liquidity ratios
Model (2): Debt ratios and *DIDC*	Model (2)*: Debt ratios
Model (3): Activity ratios and *DIDC*	Model (3)*: Activity ratios
Model (4): Profitability ratios and *DIDC*	Model (4)*: Profitability ratios
Model (5): Financial ratios and *DIDC*	Model (5)*: Financial ratios
Model (6): Significant financial ratios and *DIDC*	Model (6)*: Significant financial ratios
Model (7): Significant financial ratios, *NMDD* and *DIDC*	Model (7)*: Significant financial ratios and *NMDD*

regression analysis are twofold: (i) to derive independent variables with significant explanatory power; and (ii) to use these independent variables to predict the probability of financial distress through the constructed model. We adopt the binary logistic regression model to carry out our quantitative analysis in the present study, with the dependent binary variable being assigned a value of 1 for financially distressed firms; otherwise 0.

We apply logistic regressions in the present study to construct an early warning model for the four quarterly periods prior to the occurrence of financial distress. Adhering to the fundamental concept of logistic regression, the default prediction model with consideration of the *DIDC* variable is specified as:

$$\text{Prob}(\text{default}=1)_{i,q+j}=\alpha_0+\alpha_1 DIDC_{iq}+\beta X_{iq}+\varepsilon_{iq} \tag{6}$$

where Prob(default = 1) refers to the probability of a firm encountering financial distress; j equals 1 to 4 implying the one-, two-, three- and four-quarter-ahead prediction, *DIDC* is the proxy for the probability of default extracted from public news; β is a coefficient vector; and X is the vector of the explanatory variables, *Liquidity Ratio*, *Debt Ratio*, *Activity Ratio*, *Profitability Ratio* and default variables. We further compare the default prediction when the market-based information of the '*naïve probability of the Merton Distance to Default*' (*NMDD*) model is taken into consideration. Table 3 describes the alternative and competitive models.

3.2. Predictive accuracy of the model

The method used for the calculation of the percentage of distressed and non-distressed firms 'correctly classified' (*CC*) is illustrated in Table 4, where *C* refers to the number of Type I errors (the number of distressed firms in the sample based on actual observations of their misclassification as non-distressed firms); *B* denotes the number of Type II errors (the number of non-distressed firms in the sample based on actual observations of their misclassification as distressed firms); and *A* and *D* respectively represent the number of non-distressed and distressed firms correctly classified by the models. Based upon the accurate determination of a standard measure for the correctly-classified firms, we can then go on to determine whether the constructed model has superior predictive power.

Table 4
Check for robustness of the model.

Observed values	Predicted values		Classification accuracy
	Non-distressed firms (Y = 0)	Distressed firms (Y = 1)	
Non-distressed firms (Y = 0)	A	B	E = A/(A + B)
Distressed firms (Y = 1)	C	D	F = D/(C + D)
Overall classification accuracy			CC = (A + D)/(A + B + C + D)

Note: 'Type I' errors are those involving the misclassification of an observed distressed firm as a non-distressed firm, which is the ratio of C/(C + D). 'Type II' errors are those involving the misclassification of an observed non-distressed firm as a distressed firm, which is the ratio of B/(A + B). Classification accuracy refers to the percentage of distressed and non-distressed firms 'correctly classified' (CC).

We also adopt the 'weighted efficiency' (*WE*) measure of Korobow and Stuhr (1985) to take into account the additional costs arising from the cutoff relationship between Type I and Type II errors. The previous conventional measure, the 'percentage correctly classified' (*CC*) was potentially quite high, even when a small percentage of weak or failed banks were found to be correctly classified. Korobow and Stuhr dealt with this problem by weighting *CC* by two additional factors: (i) banks that actually weakened or failed as a percentage of those failing the 'hurdle test' of the model; and (ii) the percentage of all weak or failed banks correctly classified. The weighted efficiency of each model is calculated as follows:

$$WE = \frac{D}{(B+D)} \times \frac{D}{(C+D)} \times CC, CC = \frac{(A+D)}{(A+B+C+D)} \tag{7}$$

where *CC* is the percentage of distressed and non-distressed firms correctly classified; *D* is the number of distressed firms correctly identified by the model; $(B+D)$ refers to the distressed firms predicted by the model; and $(C+D)$ denotes the total number of distressed firms observed within the sample.

The fit of the logistic models used in this study is validated by comparing the predicted value of each observation with the cutoff value. If the predicted value is above the cutoff value, then the firm is classified as a 'distressed' firm; otherwise, it is classified as a 'non-distressed' firm. With the exception of the 0.5 probability threshold, we follow the suggestion of Martin (1977) to use the empirical cutoff value, which is the percentage of financially distressed firms in the overall study sample $(81/[81+333]=0.1956)$.

3.3. Comparing the default forecasting of the models

We employ a moving window approach with an augmenting information set to test the robustness of the predictive power of the constructed models, as shown in Fig. 1. The samples were classified into sub-groups according to the time series of the financial distress event, with the predictive power of the models then being examined for the different sub-groups. We use the values of the explanatory variables for the estimation samples in the time periods to generate the parametric estimates of each variable, and then derive the probability of financial distress for the corresponding validation samples from each of the logistic regression models.

We subsequently go on to determine whether *DIDC* improves the forecasting model through a reduction in forecasting error or an increase in classification accuracy by including only the financial variables, with the marginal contribution of *DIDC* being calculated as follows:

$$MC_{TypeI} = TypeI_{Modeli} - TypeI_{Modeli*} \tag{8}$$

$$MC_{TypeII} = TypeII_{Modeli} - TypeII_{Modeli*} \tag{9}$$

$$MC_{CC} = CC_{Modeli} - CC_{Modeli*} \tag{10}$$

$$MC_{WE} = WE_{Modeli} - WE_{Modeli*} \tag{11}$$

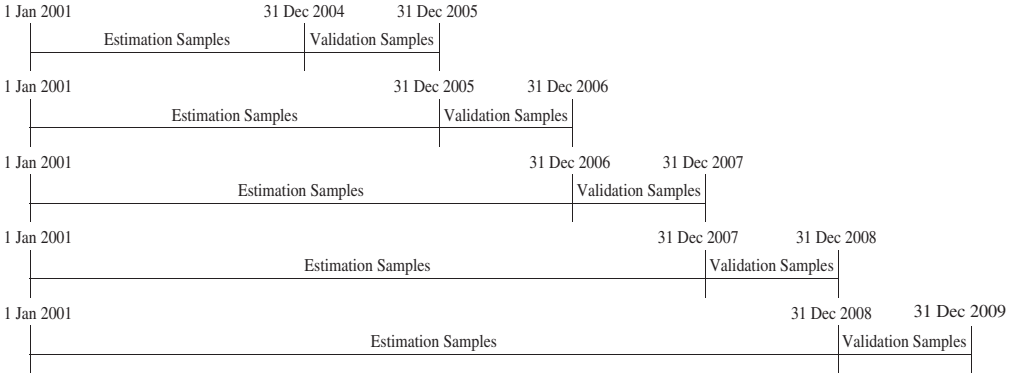


Fig. 1. Moving window of estimation and validation samples.

where MC_{TypeI} (MC_{TypeII}) denotes the marginal contribution of Type I (Type II) errors which is reduced when *DIDC* is included in the model; MC_{CC} refers to the marginal contribution of the 'correctly classified' (*CC*) samples which is increased when *DIDC* is included in the model; and MC_{WE} is the same measure for the 'weight efficiency' (*WE*) samples. Model (i) is the default forecasting model incorporating the alternative financial ratios and *DIDC*; and Model (i^*) is the default forecasting model which includes only the alternative financial ratios. If there is a reduction (increase) in MC_{TypeI} and MC_{TypeII} (MC_{CC} and MC_{WE}) then *DIDC* can be regarded as having predictive power on corporate credit default.

4. Empirical results

We begin this section by presenting the descriptive statistics of the samples and the variables, followed by the construction of our distress probability prediction model using logistic regression analysis; we then go on to analyze the empirical results. The robustness of the model is examined using classification accuracy and a moving window design with an augmenting information set.

4.1. Descriptive statistics

The total numbers and relative ratios of the distressed and non-distressed firms are presented in Table 5. Following the exclusion of firms in the banking and insurance sectors (based upon their special business nature and financial structure), our study sample ultimately comprised of 414 firms; of this total, 81 (333) were distressed (non-distressed) firms. The number of samples in Table 5 is much larger than the number of firms since the data were collected from the first, second, third and fourth quarters prior to the financial distress event.

Of the total number of firms, 19.57% were distressed firms; of the total study sample, 18.06% were distressed samples. Both of these are very close to 1:4 ratios of distressed to non-distressed groups. The ratio of distressed samples is slightly lower than the ratio of distressed firms, possibly because there may have been no public news reports on some firms prior to their default event.

The summary statistics of *DIDC* on the full sample, and the distressed and non-distressed samples, are presented in Table 6, from which we can see that the mean, median, mode, minimum, maximum and standard deviations are all greater in the distressed group than in either the full sample or the non-distressed sample. The skewness and kurtosis show that the distressed group has leptokurtic distribution, which is skewed to the right, while the non-distressed group is close to leptokurtic distribution with left skewness.

The trend in the *DIDC* prior to the occurrences of financial distress is illustrated in Fig. 2, with the distribution being subsequently illustrated in Fig. 3. Any *DIDC* outliers which are greater than 2 or smaller than 0.5 in Fig. 2 are subsequently deleted in order to facilitate further comparison of the distribution of the distressed and non-distressed firms. The mean *DIDC* of the distressed firms is found to be 1.1926, while the mean for the

Table 5
Ratio of distressed firms to non-distressed firms.

Year	No. of firms			No. of samples		
	Distressed firms (A)	Non-distressed firms (B)	Ratio (%) = (A)/[(A) + (B)]	Distressed samples (C)	Non-distressed samples (D)	Ratio (%) = (C)/[(C) + (D)]
2001	13	45	22.41	42	157	21.11
2002	6	12	33.33	8	15	34.78
2003	2	29	6.45	8	105	7.08
2004	9	30	23.08	23	108	17.56
2005	12	43	21.82	35	145	19.44
2006	4	27	12.90	10	66	13.16
2007	6	25	19.35	13	72	15.29
2008	15	71	17.44	48	201	19.28
2009	14	51	21.54	40	161	19.90
Totals	81	333	19.57	227	1030	18.06

Note: 'No. of firms' includes all distressed and non-distressed firms examined during the study period. 'No. of samples' is the *DIDC* calculated in the four quarterly periods (0–3 months, 4–6 months, 7–9 months and 10–12 months) prior to the event quarter. Given that some of the firms may not have had any media-related news during the four quarters prior to the event quarter, the number of samples is less than four times the number of firms.

Table 6
Summary statistics of *DIDC*.

Variables	Full sample	Distressed samples	Non-distressed samples
Mean	1.0811	1.1926	1.0565
Median	1.0898	1.1661	1.0776
Mode	1.0202	1.0806	1.0202
S.D.	0.1672	0.2210	0.1417
Min.	0.0000	0.6677	0.0000
Max.	3.9160	3.9160	2.4970
Skewness	2.8451	8.5687	−2.2620
Kurtosis	80.7269	103.8017	30.6189
Total no.	1257	227	1030

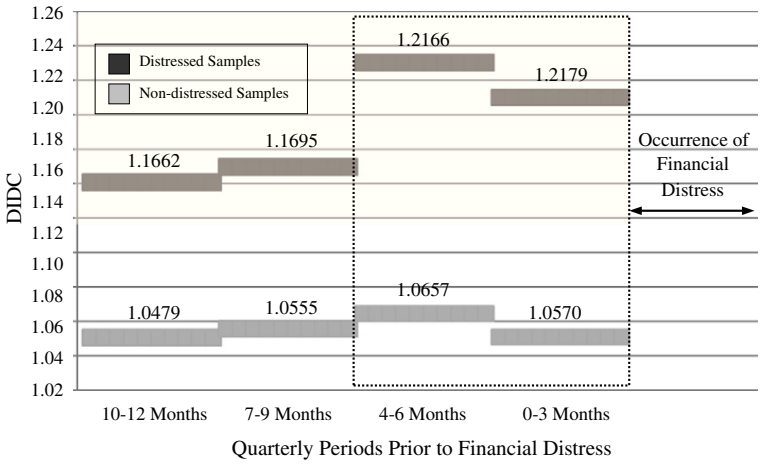
non-distressed firms is 1.0565, with these results confirming that the higher the *DIDC* indicator extracted from the financial news, the greater the likelihood of the sampled firm being found to be a distressed firm.

4.2. Mean difference tests between distressed and non-distressed firms

The variations between distressed and non-distressed firms are shown in Table 7, with Panels A to F respectively presenting the *Liquidity Ratio*, *Debt Ratio*, *Activity Ratio*, *Profitability Ratio*, *NMDD* and *DIDC* of the distressed and non-distressed firms during the four quarters prior to the occurrence of financial distress.

The *Liquidity Ratio* results in Panel A indicate that during the four quarters prior to default, the *Current Ratio*, *Quick Ratio* and *Cash Flow Ratio* are all significant at the 1% level, while *Interest Ratio* is significant at the 1% level between the third and fourth quarters. The results for *Debt Ratio* in Panel B show that during all four quarters prior to financial distress, *Debt Ratio* is significant at the 1% level and *Times Interest Earned Ratio* is significant at the 5% level. No significant difference is found between the distressed and non-distressed groups with regard to the variation in *Net Debt* except for the second quarter prior to distressed event.

The results for *Activity Ratio* in Panel C indicate that only *EPS* exhibits any consistently significant difference at the 1% level during the four quarters prior to default. The results for *Profitability Ratio* in Panel D reveal that *Profit Growth*, *Operating Profit Margin* (the other profitability measure including *Profit Growth*),



Note: The quarterly periods represent the three-month periods prior to the occurrence of a financial distress event, from the 10- to 12-month period to the 0- to 3-month period prior to the event. The numbers indicate the average 'distress intensity of default-corpus' (*DIDC*) for both the distressed and non-distressed samples.

Fig. 2. Distress intensity of default-corpus (DIDC) trend in the quarterly periods prior to the occurrence of financial distress events.

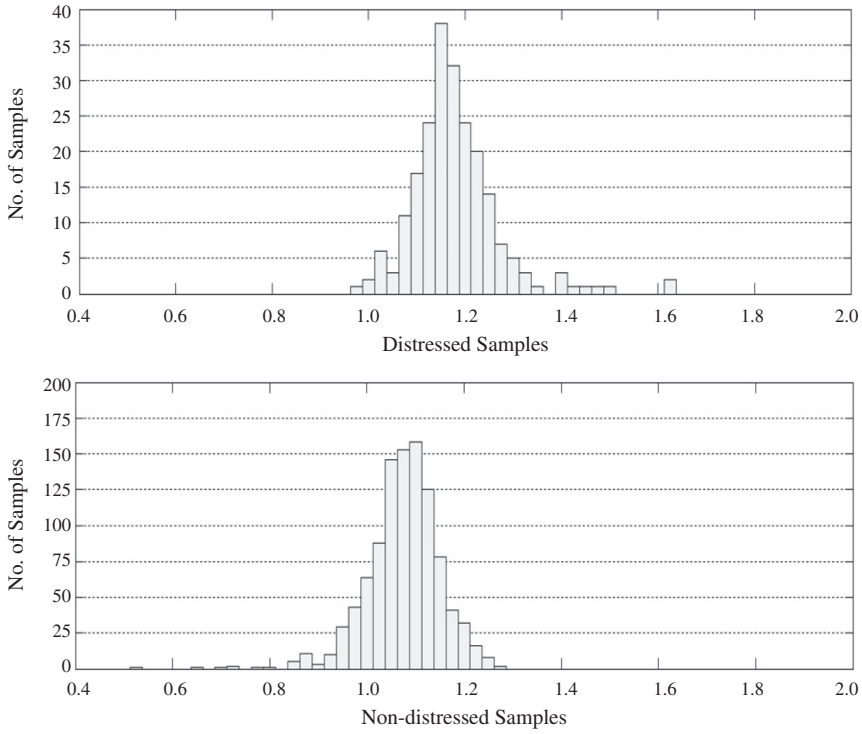


Fig. 3. Distribution of distress intensity of default-corpus (DIDC).

EBT/Sales, *ROA* and *ROE* are all significant at the 5% to 1% levels during the four quarters prior to distress. *Profit Margin* is significant at the 5% level in the first and third quarters immediately prior to financial distress.

The variations in *NMDD*, shown in Panel E of Table 7, reveal significant differences between the distressed and non-distressed samples at the 5% to 1% levels during the four quarters prior to distress. The variations in *DIDC* are shown in Panel F of Table 7, where the mean differences between distressed and non-distressed samples provide strong support for the use of default models incorporating the financial ratios and *DIDC*.

4.3. Logistic regression results

The logistic regression results for the four quarterly periods prior to financial distress are presented in Table 8, with the variables combining the significant financial ratios, *NMDD* and *DIDC*. Although we find that the explanatory power of the forecasting model is improved with the inclusion of more variables in the models, we also note that efficacy is reduced when certain meaningless variables are included. Therefore the financial ratios used in Table 8 are those exhibiting significant warning signals during the four quarters prior to financial distress.[7]

The results in Table 8 show that the *NMDD* calculated one quarter prior to the financial distress event does not have a significant correlation with default prediction; though there are significant different mean differences of *NMDD* between distress and non-distressed samples in Panel E of Table 7. Our examination indicates that the explanatory power of *NMDD* is weak when *DIDC* is incorporated. The explanatory power

[7] The regression coefficients of the logistic models incorporating *DIDC* and liquidity ratio, debt ratio, activity ratio, profitability ratio and all of the financial ratios individually are omitted for space considerations but are available from the authors upon request. The financial ratios used in Table 8 (those exhibiting significant warning signals during the four quarters prior to the financial distress event) are current ratio, quick ratio, cash flow ratio, receivable turnover ratio, interest ratio, debt ratio, total asset turnover ratio, EPS, net worth turnover ratio, EBITDA, profit growth, EBT/sales, ROA, and ROE.

Table 7
Variations between the distressed and non-distressed sample firms.

Variables	Firm type	Quarterly periods prior to financial distress											
		0–3 months			4–6 months			7–9 months			10–12 months		
		Mean	S.D.	t-value	Mean	S.D.	t-value	Mean	S.D.	t-value	Mean	S.D.	t-value
Panel A: *Liquidity Ratio*													
Current Ratio	DF	124.49	79.67	−7.28***	139.17	180.63	−3.46***	122.21	52.75	−8.53***	130.54	61.96	−6.03***
	NDF	229.61	166.97		230.98	155.66		245.95	203.83		234.69	242.12	
Quick Ratio	DF	75.54	72.78	−6.91***	81.82	172.10	−3.27***	70.41	47.24	−7.83***	81.25	54.48	−5.31***
	NDF	162.70	138.37		164.08	143.93		176.19	192.13		167.21	230.74	
Cash Flow Ratio	DF	−0.77	41.91	−2.65***	−7.74	50.24	−3.96***	−4.49	22.54	−6.82***	−7.05	27.51	−6.59***
	NDF	16.27	57.22		23.45	61.71		28.36	60.27		22.39	41.67	
Receivable Turnover Ratio	DF	6.14	16.00	−1.31	5.54	17.64	−1.31	4.63	7.38	−1.50	4.18	6.46	−0.97
	NDF	11.12	52.54		10.42	44.76		7.97	32.12		5.22	10.21	
Inventory Turnover Ratio	DF	6.82	15.61	−1.25	6.60	16.96	−0.49	6.02	10.80	−1.40	5.43	6.95	−1.71*
	NDF	11.81	56.83		7.93	21.83		9.05	25.89		8.85	28.31	
Interest Ratio	DF	13.27	61.98	1.43	17.25	84.40	1.40	4.58	8.66	2.98***	3.33	4.08	3.85***
	NDF	1.80	6.15		1.17	1.96		1.08	1.68		1.18	2.43	
Panel B: *Debt Ratio*													
Net Debt	DF	23.46	132.41	−0.85	11.87	64.91	−2.35**	17.68	70.12	−1.01	22.12	50.14	−0.22
	NDF	89.66	1251.37		82.08	447.88		34.49	222.21		33.05	779.20	
Debt Ratio	DF	61.60	19.03	8.64***	60.76	16.55	8.95***	61.08	13.98	11.00***	54.66	14.57	7.2***
	NDF	38.88	15.39		38.86	14.97		38.01	14.83		39.36	14.69	
Times Interest Earned Ratio	DF	6.65	189.35	−3.15***	11.30	163.38	−2.68***	−8.26	20.13	−2.37**	−50.74	258.86	−2.67***
	NDF	669.42	3443.15		682.22	3914.85		1110.95	7624.17		762.01	4798.09	
Panel C: *Activity Ratio*													
Total Asset Turnover Ratio	DF	0.41	0.48	−1.91*	0.38	0.39	−2.88***	0.57	0.55	−1.59	0.49	0.41	−2.13**
	NDF	0.54	0.49		0.56	0.49		0.70	0.66		0.63	0.58	

EPS	DF	−1.90	3.97	−5.78***	−1.71	2.99	−6.60***	−1.75	2.85	−7.81***	−0.65	1.62	−7.68***
	NDF	1.15	2.09		1.12	2.16		1.40	2.00		1.46	2.76	
Equity to Net Worth	DF	42.01	46.84	1.42	54.33	85.54	1.93	38.76	36.99	1.64	32.89	33.21	0.55
	NDF	33.02	31.20		31.64	28.96		30.11	27.40		30.32	27.71	
Net Worth Turnover Ratio	DF	1.51	3.28	1.25	1.00	1.07	0.15	1.65	1.94	1.44	1.22	1.16	0.54
	NDF	0.97	1.02		0.98	0.96		1.25	1.44		1.13	1.24	
EBITDA	DF	−374.43	1089.18	−2.58***	174.60	1925.39	0.58	−84.56	1327.80	−0.61	−263.51	1802.05	−1.41
	NDF	−2.34	516.58		22.99	286.71		26.81	622.07		76.55	755.66	
Panel D: *Profitability Ratio*													
Profit Growth	DF	3.73	42.76	−2.94***	9.08	19.40	−3.78***	7.31	19.15	−4.73***	7.95	13.57	−6.42***
	NDF	20.48	22.39		19.95	18.06		20.62	18.07		21.15	16.30	
Profit Margin	DF	−62.67	230.24	−1.96**	−124.25	725.78	−1.59	−14.16	174.63	−2.09**	150.42	661.04	0.89
	NDF	0.56	204.09		34.06	218.52		40.62	187.65		68.29	504.90	
Operating Profit Margin	DF	−35.88	128.83	−2.45**	−33.47	88.95	−3.19***	−11.45	32.87	−4.04***	−6.11	17.16	−5.73***
	NDF	5.00	21.82		5.30	18.75		6.82	14.60		7.59	12.63	
EBT/Sales	DF	17.03	19.42	2.99***	17.21	17.99	5.04***	12.38	14.81	3.86***	10.29	13.58	4.84***
	NDF	−53.95	388.79		−22.09	116.36		−28.44	167.40		−19.96	94.99	
ROA	DF	−8.58	15.91	−6.00***	−7.03	15.07	−5.30***	−5.76	9.80	−8.02***	−2.02	6.25	−7.85***
	NDF	3.98	6.81		4.15	7.90		5.38	7.00		5.29	6.97	
ROE	DF	−33.97	75.97	−4.03***	−24.11	40.32	−5.31***	−20.16	33.65	−6.09***	−6.13	12.35	−7.75***
	NDF	5.68	11.23		5.35	12.59		7.80	11.48		7.73	12.00	
Panel E: *Market-based variable*													
NMDD	DF	0.62	0.38	6.95***	0.40	0.37	3.29***	0.41	0.38	4.17***	0.25	0.34	2.26**
	NDF	0.23	0.34		0.21	0.32		0.17	0.31		0.13	0.28	
Panel F: *News-corpus*													
DIDC	DF	1.22	0.14	8.18***	1.22	0.39	2.83***	1.17	0.11	6.35***	1.17	0.14	5.70***
	NDF	1.06	0.12		1.07	0.13		1.06	0.17		1.05	0.14	

Note: DF refers to distressed firms and NDF refers to non-distressed firms.
* Denotes significance at the 10% level.
** Denotes significance at the 5% level.
*** Denotes significance at the 1% level.

Table 8
Results of the logistic regression with the incorporation of the significant financial variables, *NMDD* and *DIDC*.

	Quarterly periods prior to financial distress							
	0–3 months		4–6 months		7–9 months		10–12 months	
Variables	Coeff.	*p*-value	Coeff.	*p*-value	Coeff.	*p*-value	Coeff.	*p*-value
Constant	−24.8285***	<0.0000	−12.8311***	0.0013	−9.7552***	0.0019	−14.5919***	0.0019
DIDC	19.4594***	<0.0000	9.1261***	0.0050	4.8024***	0.0040	12.9780***	0.0005
Current Ratio	−0.0090	0.1130	−0.0080	0.1584	−0.0136**	0.0471	−0.0193***	0.0030
Quick Ratio	0.0148**	0.0383	0.0056	0.3449	0.0142*	0.0838	0.0157*	0.0535
Cash Flow Ratio	−0.0123	0.1028	−0.0184**	0.0312	−0.0050	0.7289	−0.0268**	0.0229
Receivable Turnover Ratio	0.0122*	0.0610	−0.0990**	0.0324	−0.0464	0.6033	0.0163	0.5493
Interest Ratio	0.0098	0.3230	0.1502**	0.0462	0.0483	0.3931	−0.0518	0.6320
Debt Ratio	0.0266	0.3146	0.0357	0.1958	0.0832**	0.0237	0.0334	0.2951
Total Asset Turnover Ratio	−6.5154***	0.0022	−2.8311	0.1515	−1.2504	0.5903	−4.6997*	0.0962
EPS	0.5682**	0.0122	−0.0079	0.9787	−0.2665	0.6932	−1.0986**	0.0366
Net Worth Turnover Ratio	2.1122**	0.0155	1.2306	0.1356	0.5351	0.5577	1.9163	0.1189
EBITDA	0.0005	0.1193	0.0011*	0.0943	0.0006	0.1356	−0.0016	0.2599
Profit Growth	0.0039	0.7130	0.0326*	0.0657	−0.0205	0.2327	−0.0414**	0.0275
EBT/Sales	0.0036	0.3619	0.0304**	0.0312	−0.0018	0.6677	−0.0049	0.3399
ROA	−0.1286	0.1168	−0.0332	0.6823	−0.2320*	0.0808	0.3671*	0.0542
ROE	−0.0733*	0.0832	−0.0178	0.7064	0.0489	0.4319	−0.0724	0.3295
NMDD	0.8257	0.2131	−0.5455	0.4423	0.3260	0.6609	−0.6155	0.5078

* Denotes significance at the 10% level.
** Denotes significance at the 5% level.
*** Denotes significance at the 1% level.

of *DIDC* on default prediction is found to persist with the incorporation of the information conveyed by the stock prices of the firms. Although the alternative financial ratios exhibit only fragmented explanatory power during the first to fourth quarters, *DIDC* consistently provides a significant early warning signal during all four quarters prior to the distress event, thereby indicating that the information content of news reports prior to the event quarter could help to resolve the difficulties involved in the forecasting of corporate credit default.

Although a comparison between the mean difference tests (Table 7) and the logistic regression models incorporating the significant financial ratios, *NMDD* and *DIDC* (Table 8) reveals significant differences in some of the financial ratios between the distressed and non-distressed groups, the explanatory power of these variables is found to be weak when DIDC is included in the model, particularly in the case of the profitability ratios.

4.4. Predictive accuracy of the model

The results on the predictive accuracy (marginal contribution) of *DIDC* on the forecasting samples are presented in Table 9 (Table 10). These results are based on the 2001–2004, 2001–2005, 2001–2006, 2001–2007 and 2001–2008 periods, which are respectively used to forecast the individual default probabilities in 2005, 2006, 2007, 2008 and 2009. The average forecasting errors and classification accuracy are then calculated using each model.

The results for the Type I and II errors, 'weighted efficiency' (*WE*) and 'correctly classified' (CC, classification accuracy) are presented in Table 9 using the probability and Martin thresholds for forecasting corporate financial distress events based upon the logistic regression model. The results include the predictive accuracy of the seven models shown in Table 3. Models (1) to (6) include the different financial ratios and *DIDC*, while Model (7) considers the additional variable *NMDD* as a check for robustness. Over the four quarters prior to the distress event, the average classification accuracy ranges from 85.43 to 89.10%, based upon a probability threshold of 0.5, which is higher than the range of 84.35 to 87.92% based upon a Martin threshold of 0.1956.

A comparison between the classification accuracy of the two approaches indicates that the probability threshold has better forecasting ability than the Martin threshold. These results are, however, reversed in a comparison between average Type I errors, where the probability threshold ranges from 46.16 to 56.19%,

Table 9
Predictive accuracy of corporate financial distress for all models.

Variables[a]	Models[b]							Average
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
Panel A: 0–3 months prior to financial distress								
1. Probability threshold = 0.5								
Type I error	46.99	47.82	45.45	51.15	48.78	40.45	42.50	46.16
Type II error	2.35	1.89	2.67	3.31	7.70	3.50	5.21	3.80
CC	90.10	90.26	90.31	89.10	85.88	90.49	87.56	89.10
WE	42.16	42.14	43.16	35.62	29.87	43.89	41.91	39.82
2. Martin threshold								
Type I error	33.65	37.12	26.41	39.74	36.28	34.74	33.33	34.47
Type II error	5.28	6.91	7.33	6.17	12.28	7.68	10.49	8.02
CC	89.84	88.33	89.90	89.18	84.11	88.32	85.73	87.92
WE	47.55	37.14	49.55	39.59	30.84	40.41	38.74	40.55
Panel B: 4–6 months prior to financial distress								
1. Probability threshold = 0.5								
Type I error	59.05	44.27	48.59	67.23	60.91	52.73	60.52	56.19
Type II error	2.40	2.33	4.15	4.48	3.54	2.90	5.05	3.55
CC	88.01	91.05	88.54	85.45	87.58	88.86	84.69	87.74
WE	29.53	44.64	32.38	24.01	29.26	33.71	24.67	31.17
2. Martin threshold								
Type I error	37.77	35.77	34.45	50.95	38.95	24.45	23.35	35.10
Type II error	7.04	18.56	12.21	10.60	10.58	7.35	13.88	11.46
CC	87.50	79.00	84.53	83.60	85.22	89.94	84.18	84.85
WE	40.90	25.37	32.79	29.14	31.17	44.81	37.92	34.59
Panel C: 7–9 months prior to financial distress								
1. Probability threshold = 0.5								
Type I error	69.02	37.27	44.77	50.08	40.76	41.59	41.14	46.38
Type II error	3.02	2.50	3.03	3.79	8.73	7.29	7.33	5.10
CC	85.19	91.46	90.02	88.23	85.55	86.62	86.05	87.59
WE	18.71	47.93	40.84	34.60	29.48	33.02	35.42	34.29
2. Martin threshold								
Type I error	45.23	30.30	25.15	30.45	27.95	24.47	37.80	31.62
Type II error	10.30	10.82	8.95	10.24	11.71	11.65	13.58	11.04
CC	83.60	85.97	88.84	86.47	85.57	86.27	81.89	85.52
WE	25.37	35.13	42.95	37.49	33.48	38.79	29.33	34.65
Panel D: 10–12 months prior to financial distress								
1. Probability threshold = 0.5								
Type I error	68.27	54.34	41.44	55.88	45.58	41.54	38.62	49.38
Type II error	2.36	7.50	6.17	4.78	9.09	8.53	9.51	6.85
CC	85.85	84.15	87.37	85.99	84.52	85.67	84.46	85.43
WE	29.73	29.63	45.75	34.92	37.23	42.74	45.71	37.96
2. Martin threshold								
Type I error	40.56	27.52	28.38	32.52	36.00	33.10	28.08	32.31
Type II error	10.66	12.40	11.20	9.95	14.63	12.04	16.04	12.42
CC	84.09	85.27	86.11	86.39	81.97	84.66	81.95	84.35
WE	37.27	41.57	50.13	48.59	31.53	42.88	38.87	41.55

Note: [a] Type I and Type II errors measure the default forecasting error; CC measures the correct classification of the default forecasting model; and WE is the weighted efficiency proposed by Korobow and Stuhr (1985), which is an additional measure for evaluating the capability of models predicting severe company weakness or failure. Details of the variables included in the liquidity ratio, debt ratio, activity ratio and profitability ratio are provided in Table 1.

[b] Model (*i*) refers to the alternative models, which include the financial ratios and *DIDC*; Model (1) includes the liquidity ratio and *DIDC*; Model (2) includes the debt ratio and *DIDC*; Model (3) includes the activity ratio and *DIDC*; Model (4) includes the profitability ratio and *DIDC*; Model (5) includes all of the financial ratios and *DIDC*; Model (6) includes the significant financial ratios and DIDC; and Model (7) includes the significant financial ratios, naïve probability of Merton distance to default (*NMDD*) and distress intensity of default-corpus (*DIDC*) as our check for robustness. All figures in the table refer to percentages (%).

while the Martin threshold ranges from 31.62 to 35.10%. The results also show that classification accuracy is greater in the 0–3 month period immediately prior to financial distress, thereby indicating that classification accuracy is higher when a default event is imminent.

Table 10
Marginal contribution of DIDC to the prediction of corporate financial distress.

Variables[a]	Models[b]								Average
	(1)–(1)*	(2)–(2)*	(3)–(3)*	(4)–(4)*	(5)–(5)*	(6)–(6)*	(7)–(7)*	(6)–(7)*	
Panel A: 0–3 months prior to financial distress									
1. Probability threshold = 0.5									
MC_{TypeI}	−27.31	−16.03	−11.99	−16.03	−12.69	−10.45	−0.83	−0.83	−12.02
MC_{TypeII}	0.50	−3.20	−1.69	−0.72	−2.40	−2.53	−1.73	−2.84	−1.83
MC_{CC}	5.03	5.69	3.64	3.59	4.89	4.34	2.25	3.09	4.06
MC_{WE}	20.18	22.19	11.56	10.72	13.62	15.85	5.45	7.15	13.34
2. Martin threshold									
MC_{TypeI}	−8.46	−0.71	−14.87	−13.65	1.92	−1.54	3.33	0.83	−4.14
MC_{TypeII}	−13.48	−8.25	−10.11	−4.14	−5.80	−8.03	−4.36	−5.13	−7.41
MC_{CC}	12.50	7.71	10.91	5.98	5.08	7.44	3.33	3.91	7.11
MC_{WE}	26.56	10.99	28.59	14.13	7.83	16.50	3.60	8.01	14.53
Panel B: 4–6 months prior to financial distress									
1. Probability threshold = 0.5									
MC_{TypeI}	−10.14	−2.14	−9.82	12.32	0.00	−4.68	8.18	−2.86	−1.14
MC_{TypeII}	1.75	−3.92	−0.04	0.02	−1.15	−0.65	−0.05	1.51	−0.31
MC_{CC}	0.48	3.85	0.96	−1.79	0.91	1.49	−1.05	−0.59	0.53
MC_{WE}	4.11	14.69	1.50	−6.09	4.07	6.68	−5.32	−1.62	2.25
2. Martin threshold									
MC_{TypeI}	0.18	14.32	3.64	15.68	−3.82	−2.00	−11.82	−8.96	0.90
MC_{TypeII}	−9.67	−0.96	−4.29	1.56	−3.73	−5.90	−5.73	−6.20	−4.36
MC_{CC}	8.34	−1.37	2.65	−3.00	3.84	5.35	6.10	5.92	3.48
MC_{WE}	17.49	−4.82	6.39	−10.08	6.28	12.10	13.96	11.72	6.63
Panel C: 7–9 months prior to financial distress									
1. Probability threshold = 0.5									
MC_{TypeI}	−5.83	−10.83	3.64	3.18	−10.98	−5.98	4.17	0.00	−2.83
MC_{TypeII}	0.60	−2.43	−1.01	0.59	−0.26	−0.75	1.86	1.67	0.03
MC_{CC}	0.50	3.74	0.02	−0.80	1.91	1.53	−2.11	−1.40	0.42
MC_{WE}	−0.51	17.93	0.99	−3.81	7.67	4.75	−6.86	−2.55	2.20
2. Martin threshold									
MC_{TypeI}	3.48	5.30	−0.15	−2.50	0.68	−7.50	7.50	−0.83	0.75
MC_{TypeII}	−4.35	−7.77	−2.54	−1.89	−1.68	−2.47	2.77	0.62	−2.16
MC_{CC}	2.95	5.25	2.31	1.98	1.55	3.36	−3.45	−0.40	1.70
MC_{WE}	4.55	6.53	7.70	5.23	3.48	10.87	−10.26	−0.91	3.40
Panel D: 10–12 months prior to financial distress									
1. Probability threshold = 0.5									
MC_{TypeI}	−9.06	−18.84	−14.06	−12.17	−5.02	−10.02	−14.62	−14.62	−12.30
MC_{TypeII}	0.03	1.41	2.41	−0.65	−3.84	−2.39	3.30	3.99	0.53
MC_{CC}	1.72	3.28	0.71	3.46	4.22	3.83	0.89	0.35	2.31
MC_{WE}	3.38	15.51	9.34	12.48	15.16	15.49	9.73	9.35	11.30
2. Martin threshold									
MC_{TypeI}	−10.47	3.18	−1.78	4.06	−4.00	−2.46	−4.00	−4.00	−2.43
MC_{TypeII}	−11.76	−17.25	−11.10	−3.29	−3.47	−3.78	−5.33	−5.33	−7.66
MC_{CC}	11.13	14.17	9.48	2.44	3.79	3.81	5.12	5.12	6.88
MC_{WE}	21.66	18.33	24.73	9.16	8.43	10.14	4.24	4.03	12.59

Note: [a] Type I and Type II errors measure the default forecasting error; MC_{TypeI} and MC_{TypeII}, which are the marginal contributions of *DIDC*, measure the percentage of Type I and Type II errors which can be reduced by *DIDC* when included in the model; MC_{CC} and MC_{WE}, which are the marginal contributions of *DIDC*, evaluate the correctness of classification (CC) and weighted efficiency (*WE*) which can be increased by *DIDC* when included in the Model, where CC measures the accuracy of classification of the default forecasting model and WE, which is the weighted efficiency proposed by Korobow and Stuhr (1985), is an additional measure for evaluating the capability of models predicting severe weakness or failure of a firm. Details of the liquidity ratio, debt ratio, activity ratio and profitability ratio variables are provided in Table 1.

[b] Model (*i**) refers to the competitive models in which only the financial ratios are included. Model (*i*) refers to the alternative models, which include the financial ratios and *DIDC*; Model (1) includes the liquidity ratio and *DIDC*; Model (2) includes the debt ratio and *DIDC*; Model (3) includes the activity ratio and *DIDC*; Model (4) includes the profitability ratio and *DIDC*; Model (5) includes all of the financial ratios and *DIDC*; Model (6) includes the significant financial ratios and *DIDC*; Model (7) includes the significant financial ratios, naïve probability of Merton distance to default (*NMDD*) and *DIDC*; and Model (7)* includes the significant financial ratios and *NMDD*. All figures in the table refer to percentages (%).

The variations in Type I and Type II errors, 'weighted efficiency' (*WE*) and 'correctly classified' (CC) firms, based upon the probability and Martin thresholds, are compared in Table 10, with the results showing that, regardless of which models are used, the forecasting model with the incorporation of *DIDC* has better predictive ability of corporate financial distress. Model (1) to Model (6) in Table 10 refers to the alternative models including the financial ratios and *DIDC*. Model (1)* to Model (6)* refers to the competitive models including only the financial ratios. When *DIDC* is incorporated into the model, the Type I and Type II errors are lower, while both *WE* and *CC* are higher especially in the 0–3 months prior to financial distress.

As a check for robustness, we compare the differences in predictive accuracy between the models with and without the inclusion of the *NMDD* variable; the results are presented in Models (7)–(7)* and (6)–(7)* in Table 10, with the comparisons showing that the default prediction model with the inclusion of the *DIDC* variable continues to reduce Type I and Type II errors and increase classification accuracy, particularly under the probability threshold, in the first quarter prior to financial distress.

In general, when incorporating the information content of media coverage into the alternative models, based upon the probability threshold, Type I errors are reduced, on average, by 1.14 to 12.30%, while Type II errors are reduced, on average, by 0.31 to 1.83% only in the first and second quarters prior to the distress event. *CC* (*WE*) is increased, on average, by 0.42 to 4.06% (2.20 to 13.34%),[8] with the predictive ability of the model on credit default generally being found to be superior in the 0–3 month period prior to financial distress, as compared to predictions over longer periods.

In summary, our comparisons between classification accuracy and weight efficiency show that the probability threshold provides greater accuracy than the Martin threshold; however, the results are reversed when Type I errors are compared, since the probability threshold exhibits higher Type I errors than the Martin threshold. Our results indicate that the reduction in Type I and II errors, and the increase in classification accuracy and weight efficiency, are greater in the 0–3 month period immediately prior to the occurrence of financial distress, thereby implying that classification accuracy is higher when a crisis is imminent.

We set out in the present study with the primary aim of verifying the importance of media coverage on the prediction of corporate credit default; and indeed, we have found that by using text mining for the selection of the appropriate variables – including financial ratio, macroeconomic, governance and market-based variables – financial institutions and regulators can exploit the information content of news to construct an effective early warning mechanism of corporate credit default.

5. Conclusions

We present a new early warning signaling model of financial distress, which, in addition to the classical early warning models constructed using the liquidity ratio, debt ratio, activity ratio and profitability ratio, incorporates a news-corpus variable based upon computational linguistic text mining analysis. We propose the term 'distress intensity from default-corpus' (*DIDC*), which is designed to capture relevant information from Chinese financial news to measure the future probability of distress for a firm.

Listed firms on the Taiwan Stock Exchange (TWSE) and the GreTai Securities Market (GTSM) which experienced financial distress between the first quarter of 2001 and the fourth quarter of 2009 are used as our experimental sample group. We construct a control sample group with a profile that is identical, or as close as possible, to the sample population to overcome the potential problem of selection bias caused by over-sampling (Zmijewski, 1984). Empirical tests are performed with our logistic model using data on the distressed group and the matching sample from the first to fourth quarters prior to the occurrence of financial distress.

We adopt a moving window approach with an augmenting information set, classifying the samples into five sub-groups in line with the time series for the financial distress event, and examine the predictive power of our early warning signaling models for the different sub-groups. Our empirical results show that, as compared to the competitive models without the default-corpus variable, those models incorporating *DIDC* have better predictive ability; indeed, the incorporation of *DIDC* into the model significantly improves its predictive power on credit distress. For our robustness analysis, we examine the 'naïve probability of the

[8] Similar results are found with the application of the Martin threshold.

Merton distance to default' (*NMDD*) model, proposed by Bharath and Shumway (2008), and find that the explanatory power of *DIDC* remains significant.

Our logistic regressions with *DIDC* for the one- to four-quarter-ahead forecasts of default probability reveal that forecasting errors are significantly reduced, while classification accuracy is increased. Finally, the results indicate that the *DIDC* variable, based upon the text mining of Chinese financial news, does indeed improve predictive ability with regard to the forecasting of financial distress.

The primary contribution of this study lies in the enhancement of early warning models of financial distress through the incorporation of information content obtained from linguistic analysis. The logistic model incorporating our news-based variable may contribute to current econometric methodology by resolving the difficulty in extracting distress information from the financial corpus. The algorithm for Chinese corpus mining proposed in this study may also be applied to the distillation of relevant information obtained from financial news reports.

References

Altman, E.I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance 23, 589–609.

Atiya, A.F., 2001. Bankruptcy prediction for credit risk using neural networks: a survey and new results. IEEE Transactions on Neural Networks 12, 929–935.

Beaver, W.H., 1966. Financial ratios as predictors of failure. Journal of Accounting Research 4, 71–111.

Berkson, J., 1944. Application of the logistic function to bio-assay. Journal of the American Statistical Association 39, 357–365.

Bharath, S.T., Shumway, T., 2008. Forecasting default with the Merton distance to default model. Review of Financial Studies 21, 1339–1369.

Chen, Y., Hu, S.Y., 2001. The controlling shareholder's personal stock loans and firm performance. Working Paper. Department of Finance, National Taiwan University.

Claessens, S., Djankov, S., Lang, L.H.P., 2000. The separation of ownership and control in East Asian corporations. Journal of Financial Economics 58, 81–112.

Deakin, E., 1972. A discriminant analysis of predictors of business failure. Journal of Accounting Research 10, 167–179.

Demers, E., Vega, V., 2011. Linguistic tone in earnings press releases: news or noise? FRB International Finance Discussion Paper No. 951.

Dimitras, A.I., Zanakis, S.H., Zopounidis, C., 1996. A survey of business failures with an emphasis on prediction methods and industrial applications. European Journal of Operational Research 90, 487–513.

Doumpos, M., Zopounidis, C., 2002. On the development of an outranking relation for ordinal classification problems: an experimental investigation of a new methodology. Optimization Methods and Software 17, 293–317.

Doumpos, M., Kosmidou, K., Baourakis, G., Zopounidis, C., 2002. Credit risk assessment using a multi-criteria hierarchical discrimination approach: a comparative analysis. European Journal of Operational Research 138, 392–412.

Fayyad, U., Uthurusamy, R., 1996. Data mining and knowledge discovery in database. Communications of the ACM 39, 24–26.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996a. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM 39, 27–34.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996b. From data mining to knowledge discovery: an overview. Advances in Knowledge Discovery and Data Mining. American Association for Artificial Intelligence, Menlo Park, CA, USA, pp. 1–36.

Frakes, W.B., Baeza-Yates, R., 1992. Information Retrieval: Data Structures and Algorithms. Prentice-Hall, Englewood Cliffs, New Jersey.

Hol, S., 2007. The influence of the business cycle on bankruptcy probability. International Federation of Operational Research Societies 14, 75–90.

Hopwood, W., McKeown, J., Mutchler, J., 1994. A reexamination of auditor versus model accuracy within the context of the going-concern opinion decision. Contemporary Accounting Research 10, 409–431.

Johnson, S., Boone, P., Breach, A., Friedman, E., 2000. Corporate governance in the Asian financial crisis 1997–98. Journal of Financial Economics 58, 141–186.

Koopman, S.J., Lucas, A., 2005. Business and default cycles for credit risk. Journal of Applied Econometrics 20, 311–323.

Korobow, L., Stuhr, D., 1985. Performance measurement of early warning models: comments on west and other weakness/failure prediction models. Journal of Banking and Finance 9, 267–273.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., 1999. Corporate ownership around the world. Journal of Finance 54, 471–517.

Lee, T.S., Yeh, Y.H., 2004. Corporate governance and financial distress: evidence from Taiwan. Corporate Governance: An International Review 12, 378–388.

Lennox, C., 1999. Identifying failing companies: a re-evaluation of the Logit, Probit and MDA approaches. Journal of Economics and Business 51, 347–364.

Lu, Y.C., Chang, S.L., 2009. Corporate governance and quality of financial information on the prediction power of financial distress of listed companies in Taiwan. International Research Journal of Finance and Economics 32, 114–138.

Manning, C.D., Schütze, H., 2002. Foundations of Statistical Natural Language Processing (2nd Edition with Corrections). MIT Press, Cambridge, London.

Martin, D., 1977. Early warning of bank failure: a logistic regression approach. Journal of Banking and Finance 1, 249–276.

Merton, R.C., 1974. On the pricing of corporate debt: the risk structure of interest rates. Journal of Finance 29, 449–470.

Ohlson, J.A., 1980. Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research 18, 109–131.

Piramuthu, S., 1999. Financial credit-risk evaluation with neural and neurofuzzy systems. European Journal of Operational Research 112, 310–321.

Platt, H.D., Platt, M.B., 1991. A note on the use of industry-relative ratios in bankruptcy prediction. Journal of Banking and Finance 15, 1183–1194.

Platt, H.D., Platt, M.B., 2002. Predicting corporate financial distress: reflections on choice-based sample bias. Journal of Economics and Finance 26, 184–199.
Platt, H.D., Platt, M.B., 2006. Understanding differences between financial distress and bankruptcy. Review of Applied Economics 2, 211–227.
Rajan, R.G., Zingales, L., 1998. Which capitalism? Lessons from the East Asian crisis. Journal of Applied Corporate Finance 11, 40–48.
Tetlock, P.C., 2007. Giving content to investor sentiment: the role of media in the stock market. Journal of Finance 62, 1139–1168.
Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S., 2008. More than words: quantifying language to measure firms' fundamentals. Journal of Finance 63, 1437–1467.
Vega, C., 2006. Stock price reaction to public and private information. Journal of Financial Economics 82, 103–133.
Westgaard, S., Wijst, N., 2001. Default probabilities in a corporate bank portfolios: a logistic model approach. European Journal of Operational Research 135, 338–349.
Yang, Y., Pedersen, J.O., 1997. A comparative study on feature selection in text categorization. Proceedings of the International Conference on Machine Learning (ICML), pp. 412–420.
Yeh, Y.H., Lee, T.S., Woidtke, T., 2001. Family control and corporate governance: evidence from Taiwan. International Review of Finance 2, 21–48.
Zmijewski, M.E., 1984. Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research 22, 59–86.