Knowledge-Based Systems 128 (2017) 139–152

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods

Petr Hajek[a,*], Roberto Henriques[b]

[a] Institute of System Engineering and Informatics, Faculty of Economics and Administration, University of Pardubice, Studentská 84, Pardubice, Czech Republic
[b] NOVA IMS, Universidade Nova de Lisboa, 1070-312, Lisboa, Portugal

ARTICLE INFO

Article history:
Received 5 December 2016
Revised 27 February 2017
Accepted 2 May 2017
Available online 6 May 2017

Keywords:
Annual reports
Financial statement fraud
Text mining
Feature selection
Machine learning

ABSTRACT

Financial statement fraud has been serious concern for investors, audit firms, government regulators, and other capital market stakeholders. Intelligent financial statement fraud detection systems have therefore been developed to support decision-making of the stakeholders. Fraudulent misrepresentation of financial statements in managerial comments has been noticed in recent studies. As such, the purpose of this study was to examine whether an improved financial fraud detection system could be developed by combining specific features derived from financial information and managerial comments in corporate annual reports. To develop this system, we employed both intelligent feature selection and classification using a wide range of machine learning methods. We found that ensemble methods outperformed the remaining methods in terms of true positive rate (fraudulent firms correctly classified as fraudulent). In contrast, Bayesian belief networks (BBN) performed best on non-fraudulent firms (true negative rate). This finding is important because interpretable "green flag" values (for which fraud is likely absent) could be derived, providing potential decision support to auditors during client selection or audit planning. We also observe that both financial statements and text in annual reports can be utilised to detect non-fraudulent firms. However, non-annual report data (analysts' forecasts of revenues and earnings) are necessary to detect fraudulent firms. This finding has important implications for selecting variables when developing early warning systems of financial statement fraud.

1. Introduction

Financial statement fraud can be defined as material omissions or misrepresentations resulting from an intentional failure to report financial information in accordance with generally accepted accounting principles [1]. As reported by the Center for Audit Quality, individuals are engaged in financial statement frauds for several reasons, including personal gain, the need to meet short-term financial expectations, and a desire to conceal bad news. Fraudulent financial statements are manipulated to be convincingly similar to non-fraudulent ones [2], and common types include improper revenue recognition, understatement/overstatement of income, expenses, assets orliabilities, and misrepresentations (or omissions) in financial statement footnotes or management discussion and analysis (MD&A), see e.g. [3] for an overview.

Over the past few decades, major firms have experienced financial statement fraud, which has had a negative impact on capital markets and a loss of shareholder value. Specifically, Beasley et al. [4] report that: (1) the average fraudulent firm's stock price dropped by 16.7% in response to the initial press disclosures of an alleged fraud; (2) 28% of fraud firms were bankrupt or liquidated within two years; (3) 47% were delisted from a national stock exchange; and (4) 62% were affected by material asset sales. Moreover, four of the ten largest bankruptcies in U.S. history were associated with major financial frauds [5]. Indeed, financial fraud may be an effective indicator of substantial financial problems that cause bankruptcy [6]. Therefore, financial statement fraud has been a serious concern for investors and other capital market stakeholders.

Although detecting fraud requires in-depth expert knowledge and is thus increasingly the responsibility of external auditors, prior research has indicated that auditors fail to detect major frauds [7]. Moreover, the credibility of audit firms has been undermined due to reported conflicts of interest [8]. Manual detection is also considered time consuming, expensive, and inaccurate [9]. Thus, accurate automated systems have become a central issue in financial statement fraud detection. The enhanced detection capability of the systems is particularly important for investors

* Corresponding author.
E-mail addresses: petr.hajek@upce.cz (P. Hajek), roberto@novaims.unl.pt (R. Henriques).

http://dx.doi.org/10.1016/j.knosys.2017.05.001
0950-7051/

(to make better informed decisions), audit firms (in conducting both client acceptance and routine audits) and government regulators (to focus their investigatory efforts better) [5,10]. Thus, intelligent detection systems of financial statement fraud have been developed to provide early warning signals (red flags) that support the decision-making processes of the stakeholders. To illustrate the growing importance of these systems, the U.S. Securities and Exchange Commission (SEC) announced the development of the Center for Risk and Quantitative Analytics, utilising quantitative data and analysis to detect, investigate and prevent misconduct that harms investors.

A considerable amount of literature has been published on intelligent financial statement fraud detection using computational intelligence methods such as neural networks [11–14], decision trees [15], support vector machines (SVM) [16,17], evolutionary algorithms [18,19], and text mining [20–22]. In addition, related types of financial frauds have been investigated such as credit card fraud [23,24], securities and commodities fraud [25] and insurance fraud [26].

Specific methods have been developed taking into account the specific characteristics of each financial fraud type (for surveys, see [9,27]). In fact, text mining approaches have been gaining significant attention in financial statement fraud detection and related financial decision-making problems [28] due to large amounts of textual data containing managerial comments and explanations. Intentional misrepresentation by content deception of these texts has been observed by several studies, indicating fraudulent misrepresentation of financial statements [29]. The analysis of managerial comments is also important because most major financial statement frauds have involved senior management, who have the opportunity, ability, and incentive to commit fraud. However, far too little attention has been paid to combining financial and linguistic data in intelligent financial statement fraud prediction [30].

This paper suggests an intelligent detection system of financial statement fraud that incorporates the advantages of (1) both financial and linguistic data, (2) automatic feature selection, and (3) accurate and interpretable classification models. Thus, the primary purpose of this study was to examine the role of financial (both publicly available financial statements and analysts' forecasts) and linguistic data. Whilst prior literature has demonstrated that it is important to discard non-informative features in order to improve the accuracy of fraud prediction [30], intelligent feature selection has been neglected in the fraud detection literature. The second aim of this study was therefore to evaluate the effect of a correlation-based filter that takes into account strong correlations among fraud predictors. Thirdly, we surveyed a wide range of machine learning methods (logistic regression, Bayesian methods, decision trees, SVM, neural networks, and ensemble methods) to establish a fraud early warning system. To be adopted in practice, this system should not only be accurate but also easy to interpret [31]. However, most of the fraud detection systems reported in the literature have been developed by maximizing only the accuracy of prediction, while neglecting the interpretability. This aspect is of particular importance since developing transparent models has become crucial especially after the recent financial crisis. Therefore, we provide two interpretable Naïve Bayes-based models in this study. A BBN with "green flag" and "red flag" values can be used to calculate the probability of financial statement fraud, while a decision table/Naïve Bayes hybrid model offers a set of decision rules to detect fraudulent/non-fraudulent firms.

In this study, we examined a sample of 311 alleged fraudulent financial reports flagged by the SEC in Accounting and Auditing Enforcement Release issued during the period 2005–2015. To obtain the class of non-fraudulent firms, we identified an industry-size matched sample of control firms.

The remainder of this paper is organised as follows. The next section reviews previous literature on intelligent financial statement fraud detection, whereas Section 3 presents financial and linguistic variables and describes our dataset. In Section 4, we develop the methodology for detecting fraudulent financial reporting, including classification methods and performance metrics. Section 5 presents the results of predictions and the sensitivity analysis of feature selection, and in Section 6 we discuss the contributions of this study. Finally, Section 7 concludes the paper and presents potential future work.

2. Intelligent detection of financial statement fraud

Previous studies have reported superior detection performance of computational intelligence methods over traditional statistical methods. Accordingly, we limit our discussion of prior literature to studies that have used intelligent financial statement fraud detection. A list of related studies is presented in Table 1, showing the method used, the data, and the resulting classification accuracy.

Regarding the input variables utilised, previous studies mostly employed financial indicators taken from financial statements. This is due to the fact that the unusual values of financial variables may indicate the need to meet targets or hide losses. This pressure/incentive to commit fraudulent practice increases the potential for fraud [12,32]. Therefore, the financial variables used in prior studies covered all aspects of firms' financial performances, such as profitability, activity, asset structure, liquidity, business situation, leverage, and market value [2,33,34]. In addition, the opportunity for fraud also increases the risk of fraud, and so non-financial variables related to the opportunity were also utilised, such as insider holdings or reinvestment ratios [12]. Recent studies have begun to concentrate on linguistic variables extracted from firm-related textual documents that may contain misleading statements. Text mining is reported highly useful for financial statement fraud because large amounts of textual data is associated with this type of fraud [9]. Humpherys et al. [35] performed a textual analysis of the MD&A section of Form 10-Ks which public companies must file annually with the SEC. Humpherys et al. [35] achieved up to 67.3% accuracy by using measures of linguistic characteristics such as lexical diversity and syntactic complexity. Sentiment analysis and part-of-speech features were used by [36], demonstrating that both positive and negative sentiment is more pronounced in fraudulent reports. Minhas and Hussain [37] compared bag-of-words and sentiment analysis to show that they perform similarly in predicting financial statement fraud. Additionally, non-verbal vocal cues have shown promising results in financial statement fraud detection [30,38]. Using common pre-processing and statistical methods, the unstructured data can be transformed into quantitative variables. This enables the subsequent use of classification methods.

The most commonly used classification methods were logistic regression, neural networks, decision trees, and SVM. Logistic regression has been utilised as a traditional statistical benchmark classifier [39]. A variety of neural networks have been proposed in the literature on financial statement fraud detection, including multilayer perceptron (MLP) [11], probabilistic neural networks [33], group method data handling [32], radial basis function neural networks [16], and growing hierarchical self-organising maps [2]. Decision trees included both single decision trees (for example, C4.5 [15] or C5.0 [40]) and the ensembles of decision trees, such as stacking [15] or bagging [41]. Evolutionary computation such as genetic algorithms [18], genetic programming [32], or firefly [19] have also been used to assist decision trees' design and training. Although neural networks and decision trees can handle nonlinear characteristics of fraud detection problems, they have been criticised for poor generalisation performance. SVM, on the other hand, provide efficient generalisation on testing data [16,41],

Table 1
Previous studies using intelligent financial statement fraud detection.

Study	Data (fraudulent/non-fraudulent)	Method (% accuracy)
[15]	41/123 Greek firms	Stacking (95.1), C4.5 (91.2), SVM (78.7)
[11]	38/38 Greek firms	BBN (90.3), MLP (80.0), ID3 (73.6)
[18]	51/339 US firms	GA (90.8)
[46]	24/124 Chinese firms	CART (92.5), LR (89.6)
[33]	199/199 Greek firms	PNN (90.2), SVM (88.4), MLP (88.4), LDA (87.8)
[20]	61/61 US firms	Text mining+SVM (82.0), text mining (75.4)
[21]	126/622 US firms	Text mining (89.5)
[34]	293/79358 US firms	LR (63.7)
[32]	101/101 Chinese firms	PNN (98.1), GP (94.1), GMDH (93.0), MLP (78.8), SVM (73.4)
[35]	101/101 US firms	Text mining+C4.5 (67.3), text mining + Naïve Bayes (67.3), text mining + SVM (65.8)
[82]	100 US firms	Text mining (83.9)
[16]	25/50 Taiwanese firms	SVM (92.0), C4.5 (84.0), RBF (82.7), MLP (82.7)
[41]	51/15934 US firms	SVM (misclassification cost 0.0025), LR (0.0026), C4.5 (0.0028), bagging (0.0028), MLP (0.0030)
[17]	Taiwanese firms (ratio 1:4)	SVM (92.0), LR (76.0)
[88]	26/26 Chinese firms	Text mining (78.1)
[89]	110/440 Chinese firms	Voting (88.9), SVM (85.5), MLP (85.1), C5.0 (78.6)
[40]	66/66 Taiwanese firms	C5.0 (85.7), LR (81.0), SVM (72.0)
[90]	47/47 Taiwanese firms	C5.0 (79.0), Rough sets (78.9), MLP (67.1)
[43]	12/45 US firms	DWD (91.2), SVM (89.5), C4.5 (82.5), MLP (75.4)
[2]	113/467 Taiwanese firms	GHSOM (type I error 11.54, type II error 19.78)
[12]	127/447 Taiwanese firms	MLP (92.8), CART (90.3), LR (88.5)
[19]	101/101 Chinese firms	Firefly (sensitivity 79.5)
[73]	138/160 Chinese firms	RF (88.0), SVM (80.18), CART (66.43), k-NN (60.11), LR (42.91)
[22]	1407/4708 reports on US firms	Text mining (83.0)
[30]	41/1531 US firms	Audio + text mining (AUC 0.81)
[42]	788/2156 US firms	LR (88.4), SVM (87.7), BBN (82.5)
[36]	180/180 US firms	Text mining (81.8)

AUC – area under the ROC curve, BBN – Bayesian belief network, CART – classification and regression tree, DWD - distance weighted discrimination, GA – genetic algorithm, GHSOM – growing hierarchical self-organizing map, GMDH – group method data handling, GP – genetic programming, k-NN – k-nearest neighbors, LDA – linear discriminant analysis, LR – logistic regression, MLP – multilayer perceptron, PNN – probabilistic neural network, RBF – radial basis function neural network, RF – random forest, and SVM – support vector machine.

although they may perform poorly on noisy data [15]. To improve the performance of SVM and other machine learning methods, Kim et al. [42] used a wrapper with particle swarm optimization algorithm for feature selection.

The number of fraud firms in the data ranged from 12 to 788 [30,42]. For U.S. firms, the SEC enforcement releases were used to find a subsample of firms committing financial statement fraud. The main limitation of this approach, however, is the focus on public companies, reducing the ability to generalise study results [13]. Most studies used a pair-wise approach, matching the number of non-fraud firms with the number of fraud firms [32,35]. The corresponding year, industry, and size have been predominantly utilised as matching criteria.

Given the differences in auditing and reporting standards between the countries investigated and periods monitored, it is difficult to compare the accuracies achieved across the studies. For U.S. firms, the accuracies ranged from 63.7% [34] to 91.2% [43], whereas the best performance achieved for Greek firms was 95.1% [15], 98.1% for Chinese firms [32], and 92.8% for Taiwanese firms [12].

Financial statement fraud detection represents a binary classification problem with four possible classification outcomes [41]: (1) true positive (a fraud firm correctly classified as a fraud firm); (2) false negative (a fraud firm incorrectly classified as a non-fraud firm); (1) true negative (a non-fraud firm correctly classified as a non-fraud firm); and (2) false positive (a non-fraud firm incorrectly classified as a fraud firm). Most previous studies have used accuracy, true positive rate, and true negative rate to determine prediction performance [9]. However, false negative and false positive classifications, as referred to by [41], are associated with different misclassification costs (MC). A higher sensitivity is preferred to higher specificity in intelligent financial statement fraud detection. However, most methods have performed significantly better at detecting legitimate transactions correctly than fraudulent ones [9,32].

3. Data

We identified instances of alleged fraudulent financial reporting by registrants of the U.S. SEC disclosed by the SEC in an Accounting and Auditing Enforcement Release (AAER). This disclosure was preceded by a significant investigation involving accounting and auditing issues. In the case of violations of SEC and federal rules, the SEC can take enforcement actions against firms. As referred to by [34], this source of fraudulent firms has several important advantages. First, it is a straightforward and consistent methodology that can be easily replicated. Thus, potential bias associated with subjective categorisation can be avoided. Second, most economically significant frauds are captured by the SEC owing to the concentration on the most important cases. In fact, the AAER captures more than three times more information-relevant events than comparative databases [44]. Third, a high level of confidence in the SEC fraudulent financial reporting is due to the fact that the SEC takes enforcement actions only in case of strong evidence of manipulation (often based on insider whistleblowers or news and analysts' reports). Finally, the percentage of observations that do not prompt SEC enforcement for financial misrepresentation is low for the AAER when compared with other databases [44]. These are the reasons why this source of fraudulent financial reporting prevailed in previous studies [29].

Our search identified 311 public companies involved in alleged instances of fraudulent financial reporting during the period 2005–2015, and thus a set of 311 annual reports were used as our sample. Corporate annual reports were considered to be ideal for our investigation because they involve corporate executives publicly discussing financial information, thereby simultaneously providing financial and linguistic information. To obtain the matched sample of non-fraudulent firms, we identified firms with the corresponding market capitalisation and industry membership. Thus, our dataset included 622 firms (311 fraudulent and 311 non-fraudulent) from a wide variety of industries such as banking and financial

Table 2
Financial variables used for financial statement fraud detection.

Abbrev.	Firm size		Asset structure
TA	total assets	FA/TA	fixed assets to total assets
R	revenues		*business situation*
	corporate reputation	R gr	growth in revenues last year
IH	shares held by mutual funds	R exp	expect. growth in revenues (next 5 years)
InH	shares held by insiders	SGAE/R	SG&A expenditures to revenues
	profitability ratios		*liquidity ratio*
NI	net income	NCWC	non-cash working capital
NI/R	net income to revenues		*leverage ratio*
NM	net margin	BD/TA	book debt to total assets
OM	operating margin	Beta	*market value ratios*
ROE	return on equity	Div/P	beta regression coefficient (3 year)
ROA	return on assets	EPS	dividends to stock price
EV/EBIT	enterprise value to earnings before interest and taxes	EPS gr	earnings per share growth in earnings per share (last 5 years)
EV/EBITDA	enterprise value to earnings before interest, taxes, depreciation and amortization	EPS exp	expected growth in EPS (next 5 years)
	activity ratios	P/E	stock price to earnings
NCWC gr	growth in NCWC	PEG	stock price to earnings to EPS growth
TA/R	total assets to revenues	PBV	price to book value ratio
Cash/R	cash to revenues	RR	reinvestment rate
NCWC/R	non-cash working capital to revenues	PS	stock price to sales

services (74 firms), medical supplies (36 firms), computer software (30 firms), pharmaceuticals (26 firms), computers/peripherals (26 firms), telecommunication equipment (24 firms), machinery (24 firms), etc.

3.1. Financial variables

Financial variables were selected to identify financial irregularities, thus indicating statement fraud. As reported above, the chosen set of financial variables should cover all aspects of a firm's financial performance in order to detect diverse types of financial statement fraud, such as understating/overstating revenue, assets, expenses or liabilities. Previous literature has provided a strong theoretical evidence for the use of financial variables [5]. Our selection of the financial variables in Table 2 was therefore influenced by previous financial fraud detection studies. The input variables can be divided into nine categories: firm size, corporate reputation, profitability ratios, activity ratios, asset structure, business situation, liquidity ratios, leverage ratios, and market value ratios.

Throckmorton et al. [30] report that large firms with poor financial performance are more susceptible to financial restatements. By contrast, Persons [45] suggests that fraudulent firms are, on average, smaller than non-fraudulent firms. As a proxy for firm size, we used total assets and revenues in accordance with previous studies [30,46]. The reputation of a company is difficult to measure, although to a certain extent this factor can be inferred from information about insiders' and institutional holdings [47]. According to agency theory, increased insiders' holdings should mitigate a manager's propensity to commit fraud. However, the propensity is also strongly affected by both the gain from the fraud and the corresponding penalty [48]. Insiders' and institutional holdings may also indicate insufficient board oversight and, thus, opportunity to commit fraud [12].

Poor financial performance was found to be associated with a greater motivation to engage in management fraud [13]. Profitability ratios are considered to be the most important financial performance measures, indicating the influence of asset management, financing, and liquidity on the profit of a company. Low profitability ratios have been reported to incentivise management to overstate revenues or understate expenses [34,45]. Moreover, overstated revenues inflate the value of the operating margin [5,20]. To examine the possibility of revenue overstatement further, we included activity ratios that measure the effectiveness of asset management and indicate management's competitive ability, which may provide an incentive to overstate revenues [45,46]. In addition, management may manipulate both accounts receivable and inventories, increasing non-cash working capital [5,13,20,45]. To become more attractive for investors, firms may also be susceptible to manipulating fixed assets, for example by using inappropriate depreciation methods [46]. Another motivation for financial statement fraud is represented by the need for continued growth [30]. Previous studies suggest that sales of fraudulent firms are generally increasing [5,33]. In addition, a low ratio of selling, general, and administrative expenditures (SG&A) to revenues was reported for firms engaged in revenue fraud [5]. A low liquidity also provides an incentive for managers to commit financial statement fraud. In agreement with [45], we used non-cash working capital to measure liquidity. Leverage ratio (book debt to total assets) was used to detect whether firms were fictitiously including assets on the balance sheet without any corresponding debt [5,20]. A higher leverage provides incentives to boost financial performance as it shifts the risk from equity owner to debt owner [32].

The final category of financial variables is related to stock market incentives. Managers are particularly concerned with a high stock price because management compensation is often tied to stock price performance [34]. Market value ratios reflect how the past activity of a firm and its future outlook are perceived by the market [47]. To achieve positive perception, managers can be tempted to manipulate earnings and dividends in particular. Moreover, Throckmorton et al. [30] report that a highly volatile stock return makes firms more susceptible to financial restatements. Here we used the Beta coefficient to measure stock price volatility. Finally, price-to-book value ratio was used to evaluate growth expectations, and was found to be an unusually high indicator of fraudulent firms [34]. Similarly, a strong stock return performance was observed in the years prior to financial statement fraud [33,34].

Taken together, the dataset contained 32 input financial variables obtained mostly from the Reuters Global Market Data (http://www.reuters.com/finance/global-market-data) in the year prior to financial statement fraud. Forward EPS and PEG ratio were drawn from the Value Line database [49].

3.2. Linguistic variables

Regarding linguistic variables, previous literature has been focused on the MD&A section of annual reports because it provides

Table 3
Linguistic variables used for financial statement fraud detection.

Abbrev.	Linguistic variable
Pos	frequency count of positive words/length of MD&A
Neg	frequency count of negative words/length of MD&A
Tone	(frequency count of positive words – frequency count of negative words)/(frequency count of positive words + frequency count of negative words)
Uncert	frequency count of uncertain words/length of MD&A
Litig	frequency count of litigious words/length of MD&A
ModStrong	frequency count of modal strong words / length of MD&A
ModWeak	frequency count of modal weak words/length of MD&A
Constr	frequency count of constraining words/length of MD&A

investors with the opportunity to receive superior qualitative information on a firm's performance and prospects from the manager's perspective [22,35,36,50]. In conformity with these studies, we extracted linguistic variables from the MD&A, representing the most important textual section from the downloaded 10-Ks (Table 3).

Previous literature has shown that fraudsters are more likely to use negative and uncertain words [30,51,52]. This has theoretical foundations in an individual's nonverbal behaviour during deception [53], suggesting that deceivers often include statements indicating aversion or negative mood. Similarly, the lack of embracement suggests that deceivers lack conviction, which results in a lower degree of certainty in their statements [54]. Regarding sentiment categories, Goel and Uzuner [36] demonstrated that fraudulent annual reports on average contain three times more positive sentiment and four times more negative sentiment compared with truthful reports. Therefore, we measured the proportion of positive and negative words in the MD&A section of annual reports. In our study, the lack of conviction was represented by the proportion of uncertain words. We also included the proportion of the following important word categories developed specifically for the financial domain [50,55]: litigious, weak modal, strong modal, and constraining words. The use of word categories based on financial context knowledge helps to reduce ambiguity associated with various meanings and tones of words in individual domains. Loughran and McDonald [50] created extensive word lists of 354 positive, 2329 negative, 291 uncertain, 871 litigious, 19 modal strong, 27 modal weak, and 184 constraining words.

One major challenge of the sentiment word lists is the detection of positive words in negative statements. To address the issue of negations, we performed a collocation analysis with positive words to detect negation words such as 'no' or 'not' occurring near a positive word (within three words) [50]. The frequency of positive words was then calculated as the positive word count minus the count for negation. Another central problem is the choice of an appropriate term-weighting scheme. Following Throckmorton et al. [30] and Goel and Uzuner [36], we used the frequency counts of word categories, which is the raw term frequency where all terms are considered equally important. However, this scheme does not consider the length of the document, and so we normalised the word category counts by the length of the MD&A. Finally, we also included the overall tone, overall tone = (Pos – Neg)/ (Pos + Neg) [56].

3.3. Descriptive statistics of the data

Our study encompasses 622 U.S. firms listed on the New York Stock Exchange (NYSE) (47% of the firms) or Nasdaq (53%). To obtain the annual reports of the firms, we downloaded all 10-Ks for the firms from the EDGAR system (www.sec.gov/edgar.shtml) for the period one year prior to financial statement fraud. Note that the 10-K filing date is usually within 90 days after the end of the firm's fiscal year. In addition to financial information obtained from the 10-Ks, we collected the data for the financial variables from the publicly available Reuters Global Market Data and Value Line database. For the imputation of the missing values, we used ε-SVR (support vector regression) where all variables except the missing one were used to estimate the missing value. The ε-SVR model was trained by using radial basis function (RBF) kernel function with the radius of 0.4 and penalty parameter $C = 8$. Table 4 shows basic descriptive statistics of the sample. The difference in financial variables between fraudulent and non-fraudulent firms are largely in agreement with the results of the above-mentioned related studies. The most striking result to emerge from the data is that fraudulent firms experienced a low profitability and an unusually high earnings per share (EPS) growth and dividend yield, respectively.

Table 5 shows that the negative sentiment in annual reports was slightly higher for fraudulent firms when compared with non-fraudulent ones. As a result, the overall tone was also more negative for the fraudulent firms. These findings suggest that fraudulent firms are more likely to use negative words. However, the remaining word categories show that the fraudulent firms attempted to use language similar to that of the non-fraudulent firms.

4. Experimental setting

Fig. 1 presents the experimental setting used in this study. First, we randomly created 30 stratified samples from the dataset (composed of financial and linguistic variables) in which a 3:1 ratio was followed for training (466 firms in total, 233 fraudulent and 233 non-fraudulent) and testing data (156 firms in total, 78 fraudulent and 78 non-fraudulent). In previous studies, this ratio has usually been 9:1 with 10 [35] to 40 [30] different partitions. Since the ratio is largely dependent on the volume of data available, we could afford to use a larger part as testing data.

Second, we performed feature selection on each data partition to reduce its dimensionality. Feature selection represents an important pre-processing step in classification tasks [57]. The idea of feature selection algorithms is based on searching through all possible combinations of features in the data to find the best subset of features in terms of an evaluator (e.g. accuracy). This can be achieved by combining an attribute subset evaluator and a search method. In this paper, we used a correlation-based feature subset selection filter as evaluator and BestFirst as search method, respectively. The correlation-based filter evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. The goal is to find subsets of features with low inter-correlation and high correlation with the class [58]. BestFirst searches the space of feature subsets by greedy hill-climbing augmented with a backtracking facility [59]. We used this feature selection method for two main reasons. First, filters operate independently of any learning algorithm, and so the selected subsets of variables can be used to compare the performance of various classification algorithms. Second, the correlation-based filter works best when input variables are strongly inter-correlated. In our dataset, we observed significant correlations (at $P < 0.05$) particularly within the categories of financial and linguistic variables.

Then, we performed fraud classification using fourteen machine learning techniques including Logistic regression [60], Bayesian classifiers (Naïve Bayes (NB) [61], BBN [62], and Decision Table/Naïve Bayes hybrid classifier [63]), SVM [64], decision trees (C4.5 [65], Simple CART [70], JRip [71], and logistic model trees [72]), neural networks (MLP [66] and voted perceptron [67]), and ensemble methods (Bagging [68], Random Forests [69], and AdaboostM1 [70]). The performance of the methods was evaluated using various classification measures and compared using traditional statistical tests. We also performed a comparative analysis of the datasets with all variables, including those selected using

Table 4
Descriptive statistics (Mean ± St.Dev.) on financial variables.

	Fraud	Non-fraud		Fraud	Non-fraud
TA	55718 ± 248526	32199 ± 181387	FA/TA	0.268 ± 0.223	0.251 ± 0.211
R	31797 ± 140578	17418 ± 97138	R gr	0.089 ± 0.256	0.110 ± 0.238
IH	0.580 ± 0.292	0.617 ± 0.293	R exp	0.078 ± 0.059	0.094 ± 0.083
InH	0.249 ± 0.154	0.297 ± 0.192	SGAE/R	0.590 ± 1.930	0.691 ± 4.115
NI	1109 ± 3589	635 ± 3001	NCWC	2634 ± 24048	1261 ± 11637
NI/R	1.873 ± 30.958	−0.576 ± 9.628	BD/TA	0.344 ± 0.238	0.300 ± 0.246
NM	−1.086 ± 9.160	−0.116 ± 2.050	Beta	1.23 ± 0.71	1.07 ± 0.39
OM	−0.513 ± 3.323	−0.049 ± 2.620	Div/P	0.010 ± 0.020	0.009 ± 0.027
ROE	0.02 ± 0.64	2.20 ± 36.47	EPS	2.37 ± 1.69	2.37 ± 1.71
ROA	0.11 ± 0.93	0.15 ± 0.80	EPS gr	0.197 ± 0.139	0.075 ± 0.140
EV/EBIT	51.6 ± 90.2	58.6 ± 201.0	EPS exp	0.140 ± 0.074	0.172 ± 0.099
EV/EBITDA	45.5 ± 126.0	32.0 ± 64.0	P/E	36.5 ± 250.3	60.4 ± 439.9
NCWC gr	−472 ± 4712	−286 ± 3970	PEG	2.27 ± 3.46	2.31 ± 4.93
TA/R	3.40 ± 12.66	3.71 ± 21.89	PBV	3.11 ± 4.52	4.57 ± 14.66
Cash/R	0.614 ± 2.432	0.555 ± 1.036	RR	0.28 ± 17.49	0.48 ± 7.96
NCWC/R	−0.396 ± 12.226	0.047 ± 0.380	PS	3.24 ± 12.63	3.50 ± 22.28

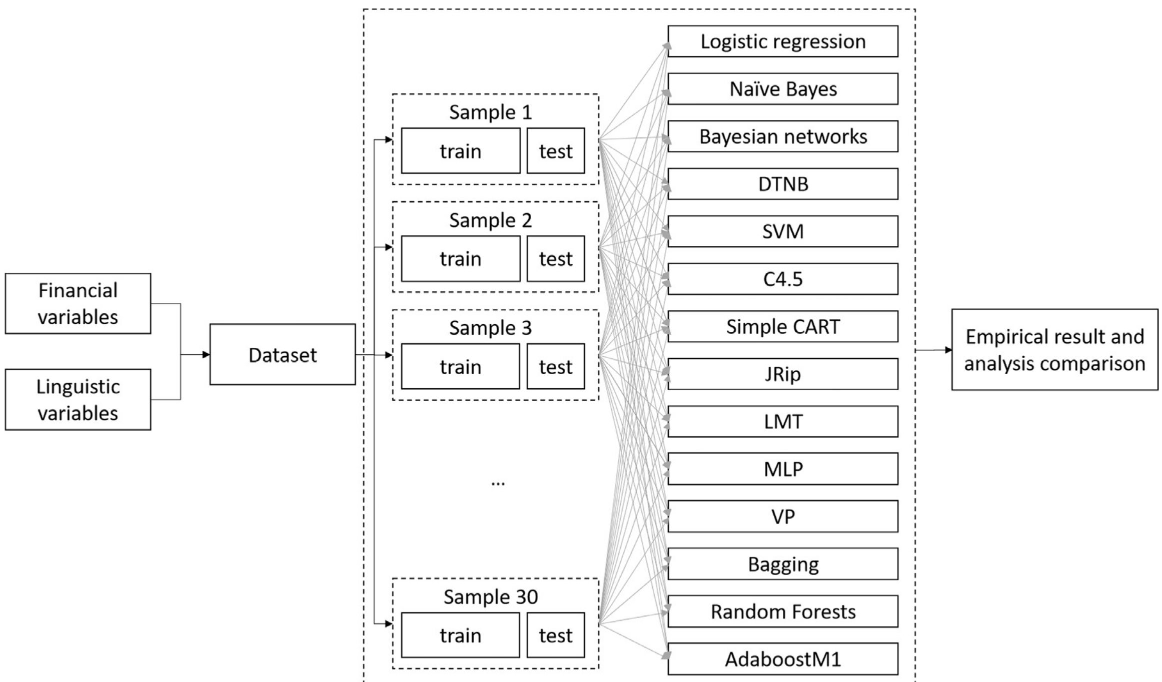

Fig. 1. Experimental setting.

Table 5
Descriptive statistics (Mean ± St.Dev.) on linguistic variables.

	Fraud	Non-fraud
Pos	0.0067 ± 0.0014	0.0068 ± 0.0015
Neg	0.0199 ± 0.0046	0.0176 ± 0.0041
Tone	−0.4826 ± 0.1294	−0.4318 ± 0.1287
Uncert	0.0134 ± 0.0027	0.0134 ± 0.0029
Litig	0.0155 ± 0.0069	0.0151 ± 0.0081
ModStrong	0.0032 ± 0.0012	0.0031 ± 0.0011
ModWeak	0.0056 ± 0.0017	0.0056 ± 0.0018
Constr	0.0074 ± 0.0014	0.0075 ± 0.0015

the correlation-based filter. Finally, we conducted sensitivity analysis of the results on the presence of linguistic variables.

4.1. Classification methods

The methods used have varied backgrounds and different theories to support them. In this way, we ensured that the problem at hand was analysed by disparate models with varying degrees of computational complexity and performance on financial fraud detection problems (see Table 1). The methods tested in this paper can be grouped into several families:logistic regression, Bayesian classifiers, support vector machines, decision trees, neural networks and ensemble classifiers.

Logistic regression is the regression analysis to conduct when the dependent variable is binary [60]. It is a discriminative classifier that is linear in its parameters, and is used to explain the relationship between one dependent binary variable and one or more metric (interval or ratio scale) independent variable. Logistic regression [60] was used in this paper in a similar way to [31] transforming the dependent variable probability into a totally continuous one.

From the Bayesian classifiers we used the NB and BBNs [62] to predict financial fraud in a similar way to [10,32]. BBNs provide a natural way to handle missing data, allow the combination of data with domain knowledge, facilitate knowledge about causal relationships between variables and provide a method for avoiding data overfitting [71]. In addition, BBNs allow for the explicit representation of dependencies using a joint probability distribution. In a similar fashion, the decision table represents a conditional probability table in the Decision Table/Naïve Bayes (DTNB) hybrid classifier [63]. This group of classifiers is not only highly computational efficient but it also provides easy to understand models.

In the SVM family, the sequential minimal optimization algorithm (SMO), which is considered state-of-the-art for nonlinear SVM, has been used for training SVMs in prior fraud detection literature [14,30]. SVM uses nonlinear mapping to transform the original training data into a higher dimension, where data from two classes can always be separated by a hyperplane. In contrast to other nonlinear methods used in this study, SVM aims to minimize structural risk, rather than empirical risk. As a result, SVM is more prone to overfitting and performs well on noisy financial fraud data [16].

Decision trees are flowcharts-like tree structure, where nodes represent attributes and branches denote possible attribute values. The attributes are chosen in terms of the goodness of a split and the sample is divided into subsets, until all the training data are correctly classified. The biggest advantage of decision trees is their interpretability of the rules generated from the model. From this family of methods, we employed J48 (an open source implementation of the well-known C4.5 algorithm), Simple CART [70], JRip (RIPPER) [71], and logistic model trees (LMT) [72]. In related fraud detection literature, decision trees performed particularly well in terms of type II error [40].

From the neural network family we employed MLP [72] and VP [67]. MLP is a multi-layer, fully connected, feed-forward network with a back-propagation training rule. Freund and Schapire [67] proposed the voted perceptron, as it is a simple algorithm with easy implementation that combines neural network and SVM algorithms. Compared with SVM, this algorithm has better precision values and learning times. Similarly to logistic regression, neural networks have a well-established history with financial fraud detection [9].

From the ensemble classifiers family we employed Bagging [68], Random Forests (RF) [69], and AdaboostM1 [70]. The ensemble models seek to extract the individual benefits of each classifier, combining them in order to get a better solution. Bagging (Bootstrap Aggregating) generates various samples with replacement from the training set. For base learners in Bagging, we used fast decision trees (REP trees) that build decision trees using information gain and reduced-error pruning. In the RF, Bagging is combined with randomly chosen subsets of input variables, which allows for a larger amount of variance reduction. Thus, the RF performed especially well in case of high-dimensional financial fraud data [73]. AdaboostM1 repeatedly runs a given weak classifier (SVM was used in this study in agreement with [74]) on various distributions over the training data, and then combines the classifiers into a single composite classifier.

The classifiers were trained using the settings presented in Table 6.

4.2. Performance metrics

Measuring the success of machine learning algorithms is an important step in determining their suitability. Classification performance can be measured in many different ways: absolute ability, performance relative to other factors, probability of success, and others [75]. In this paper we use Accuracy, *TP* rate, *TN* rate, MC (combination of *FP* and *FN* rates), *F*-measure and area under the receiver operating characteristic curve (AUC).

Accuracy is defined as the percentage of correctly classified samples

$$\text{Accuracy} = \frac{(TP + TN)}{P + N}, \tag{1}$$

where TP is the number of fraudulent companies classified as fraudulent, *TN* is the number of non-fraudulent companies classified as non-fraudulent, *P* is the number of companies classified as fraudulent, and *N* is the number of companies classified as non-fraudulent.

TP rate (also called sensitivity and recall) is the number of companies correctly classified as fraudulent as a percentage of all fraudulent companies

$$TP \text{ rate} = \frac{TP}{P}. \tag{2}$$

FP rate (type I error) is the number of companies incorrectly classified as fraudulent as a percentage of all non-fraudulent companies

$$FP \text{ rate} = \frac{FP}{N} = 1 - TN\ rate. \tag{3}$$

TN rate (also called specificity) is the number of companies correctly classified as non-fraudulent as a percentage of all non-fraudulent companies

$$TN \text{ rate} = \frac{TP}{N}. \tag{4}$$

FN rate (type II error) is the number of companies incorrectly classified as non-fraudulent as a percentage of all fraudulent companies

$$FN \text{ rate} = \frac{FN}{P} = 1 - TP \text{ rate}. \tag{5}$$

As indicated above, *FN* and *FP* rates incur substantially different MC. In fact, the costs of failing to detect financial statement fraud (*FN* rate) is much higher than predicting fraud when it is not present (*FP* rate) [76]. Note that this may seem in contradiction with related business failure classification where the cost of classifying a firm that failed as solvent (*FP* costs) is much higher than classifying as insolvent a firm that does not fail (*FN* costs) [77]. The difference is because fraud is usually considered a positive class, whereas business failure is taken as a negative class. In other words, fraud and failure are both associated with increased MC. To estimate the *FN* costs, we used the median loss of $975,000 attributable to financial statement fraud as calculated for the years 2014–2015 by the Association of Certified Fraud Examiners [78], while median audit fees ($522,205) [79] were used for *FP* costs. Therefore, we used MC with the approximate cost ratio of 1:2 as follows

$$\text{MC} = FP \text{ rate} + 2 \times FN \text{ rate}. \tag{6}$$

F-measure, also known as *F*-score or *F*, is the harmonic mean of precision and *TP* rate

$$F - \text{measure} = 2 \times \frac{\text{Precision} \times TP \text{ rate}}{\text{Precision} + TP \text{ rate}}. \tag{7}$$

AUC is equivalent to the probability that the classification model will rank a randomly chosen positive instance higher than a randomly chosen negative instance [80]. AUC was reported preferable for financial fraud detection by [30] owing to the robustness to imbalanced data.

5. Experimental results

In the first set of experiments, we used the original datasets with all variables (i.e. without feature selection) to demonstrate both the effect of feature selection and the effect of financial and linguistic variables. Table 7 summarises the results of the experiments for all variables. For all classification metrics, average values from the 30 runs are presented along with standard deviations. In addition, the best and statistically similar results (at $P < 0.05$ using a Student's paired *t*-test) are marked with asterisks. The results

Table 6
Settings of machine learning methods.

Method	Parameters and their values
Logistic regression	Broyden–Fletcher–Goldfarb–Shanno learning algorithm
NB	no kernel estimator
BBN	K2 (a hill climbing algorithm restricted by an order on the variables) with no. of parents = {1, 2} and Bayes scoring function
DTNB	no. of folds for cross validation = leave one out, search method used to find good attribute combinations = BackwardsWithDelete
SVM	SMO algorithm with complexity parameter $C = \{2^0, 2^1, 2^2, \ldots, 2^5\}$, polynomial kernel function with exponent = {1, 2}, RBF kernel function with gamma = 0.01
JRIP	no. of folds = 3, and no. of optimizations = 2
C4.5	J48 version with the minimum no. of instances per leaf = 2, and confidence factor for pruning = 0.25
CART	no. of folds in the internal cross-validation = 5, minimal no. of observations at the terminal nodes = 2
LMT	minimum no. of instances at which a node is considered for splitting = 15
MLP	neurons in hidden layer = $\{2^2, 2^3, 2^4, 2^5\}$, learning rate = 0.1, and no. of iterations = 500
VP	exponent for the polynomial kernel = 1, the maximum no. of alternations to the perceptron = 1000, and no. of iterations = 1
Bagging	base learner = REP tree, no. of iterations = 10, bag size as a percentage of the training set = {25,50,100}
RF	maximum depth of trees unlimited, no. of trees to be generated = 100, and no. of variables randomly sampled as candidates at each split = $\log_2$(#predictors) + 1
AdaboostM1	classifier = SMO algorithm with complexity parameter $C = \{2^0, 2^1, 2^2, \ldots, 2^5\}$, polynomial kernel function with exponent = {1, 2}, RBF kernel function with gamma = 0.01, and no. of iterations = 10

Table 7
Classification performance (mean ± std.) for 30 testing datasets using all variables.

	Accuracy	TP rate	TN rate	F-measure	AUC	MC
LR	74.53 ± 2.96	72.09 ± 4.54	77.09 ± 5.83	0.745 ± 0.030	0.791 ± 0.031	0.787 ± 0.149
NB	57.83 ± 3.61	28.06 ± 16.45	86.99 ± 12.89	0.526 ± 0.056	0.700 ± 0.040	1.569 ± 0.458
BBN	90.32 ± 1.90*	85.19 ± 4.26*	95.39 ± 2.51*	0.903 ± 0.019*	0.979 ± 0.007*	0.342 ± 0.110*
DTNB	89.50 ± 1.91*	87.21 ± 3.24*	91.89 ± 3.14	0.895 ± 0.019*	0.974 ± 0.008*	0.337 ± 0.096*
SVM	77.95 ± 3.05	76.66 ± 5.12	79.41 ± 4.68	0.779 ± 0.031	0.780 ± 0.031	0.673 ± 0.149
JRIP	87.01 ± 2.90	86.46 ± 4.56*	87.74 ± 6.91	0.870 ± 0.029	0.904 ± 0.023	0.393 ± 0.160
C4.5	86.10 ± 2.01	86.43 ± 4.68*	85.89 ± 4.99	0.861 ± 0.020	0.876 ± 0.041	0.412 ± 0.143
CART	86.24 ± 2.51	86.10 ± 4.01*	86.45 ± 4.04	0.862 ± 0.025	0.919 ± 0.037	0.413 ± 0.121
LMT	85.44 ± 2.33	85.33 ± 3.48*	85.56 ± 4.03	0.854 ± 0.023	0.912 ± 0.026	0.438 ± 0.110
MLP	77.93 ± 3.23	76.57 ± 5.93	79.53 ± 6.03	0.779 ± 0.033	0.848 ± 0.029	0.673 ± 0.179
VP	51.16 ± 3.94	83.21 ± 12.74	19.83 ± 16.37	0.440 ± 0.083	0.518 ± 0.046	1.138 ± 0.419
Bag	87.09 ± 2.42	86.12 ± 3.50*	88.15 ± 4.56	0.871 ± 0.024	0.959 ± 0.012	0.396 ± 0.116
RF	87.50 ± 2.35	86.93 ± 4.27*	88.13 ± 3.64	0.875 ± 0.024	0.965 ± 0.010	0.380 ± 0.122
AB	77.29 ± 3.06	75.59 ± 5.46	79.16 ± 4.36	0.773 ± 0.031	0.798 ± 0.033	0.697 ± 0.153

* best and statistically similar classification performance at $P<0.05$ using Student's paired t-test

Table 8
Classification performance (mean ± std.) for 30 testing datasets using financial variables.

	Accuracy	TP rate	TN rate	F-measure	AUC	MC
LR	71.98 ± 2.67#	69.42 ± 3.26#	74.67 ± 4.96	0.720 ± 0.027#	0.762 ± 0.028#	0.865 ± 0.115#
NB	57.08 ± 3.15	26.78 ± 15.62	86.78 ± 13.35	0.517 ± 0.050	0.685 ± 0.038	1.597 ± 0.446
BBN	90.09 ± 1.80	84.99 ± 3.76	95.13 ± 2.70	0.901 ± 0.018	0.978 ± 0.007	0.349 ± 0.102
DTNB	89.70 ± 1.99	87.36 ± 3.50	92.12 ± 3.14	0.897 ± 0.020	0.975 ± 0.008	0.332 ± 0.101
SVM	77.36 ± 3.11	79.95 ± 4.80*	74.95 ± 5.80#	0.773 ± 0.031	0.774 ± 0.031	0.652 ± 0.154
JRIP	87.15 ± 2.89	85.97 ± 5.17	88.67 ± 7.57	0.871 ± 0.029	0.898 ± 0.027	0.394 ± 0.179
C4.5	85.83 ± 2.00	85.23 ± 4.73	86.47 ± 4.62	0.858 ± 0.020	0.871 ± 0.039	0.431 ± 0.141
CART	86.36 ± 3.11	86.70 ± 4.36	86.16 ± 5.87	0.864 ± 0.031	0.925 ± 0.031	0.404 ± 0.146
LMT	85.06 ± 2.26	84.59 ± 4.13	85.56 ± 4.53	0.850 ± 0.023	0.915 ± 0.023	0.453 ± 0.128
MLP	79.48 ± 2.47*	77.28 ± 7.31	81.89 ± 5.51	0.794 ± 0.025*	0.870 ± 0.023*	0.635 ± 0.201
VP	51.07 ± 4.32	82.07 ± 13.74	20.89 ± 17.03	0.441 ± 0.084	0.518 ± 0.046	1.150 ± 0.445
Bag	87.13 ± 1.78	86.82 ± 4.10	87.50 ± 3.76	0.871 ± 0.018	0.960 ± 0.009	0.389 ± 0.120
RF	88.07 ± 2.13	87.75 ± 4.33	88.51 ± 3.91	0.881 ± 0.021	0.965 ± 0.010	0.360 ± 0.126
AB	77.01 ± 3.41	79.27 ± 4.88*	74.91 ± 5.77#	0.770 ± 0.034	0.796 ± 0.031	0.666 ± 0.155

* statistically better and # statistically worse classification performance than for all variables at $P<0.05$ using Student's paired t-test

show that BBN and DTNB significantly outperformed the remaining methods in terms of most classification metrics. On one hand, the BBN performed best in terms of *TN* rate. On the other hand, the DTNB performed best in terms of *TP* rate, leading to the lowest MC. In contrast, LR, VP and NB performed poorly in the classification of the fraud class.

In the second set of experiments, we used all the financial variables. The results in Table 8 suggest that most classifiers (except LR) performed statistically similar for all variables. In other words, omitting the linguistic variables from the set of input variables did not deteriorate the performance of the classifiers. To examine the role of the linguistic variables further, we performed the third set of experiments with the linguistic variables only (Table 9). As can be seen from Table 9, the classifiers performed significantly worse without financial variables, which indicates that predicting financial fraud using the linguistic variables only leads to relatively poor classification performance, regardless of the classifier.

Table 9
Classification performance (mean ± std.) for 30 testing datasets using linguistic variables.

	Accuracy	TP rate	TN rate	F-measure	AUC	MC
LR	61.93 ± 3.46#	57.58 ± 4.82#	66.43 ± 6.50#	0.618 ± 0.034#	0.670 ± 0.031#	1.184 ± 0.161#
NB	61.94 ± 2.89#	58.82 ± 4.66#	65.17 ± 5.55#	0.619 ± 0.029*	0.647 ± 0.034#	1.172 ± 0.149*
BBN	58.32 ± 3.15#	52.43 ± 14.93#	64.85 ± 15.13#	0.571 ± 0.055#	0.600 ± 0.031#	1.303 ± 0.450#
DTNB	58.15 ± 3.61#	49.16 ± 15.25#	67.86 ± 15.09#	0.567 ± 0.059#	0.594 ± 0.030#	1.338 ± 0.456#
SVM	62.48 ± 2.83#	56.53 ± 4.55#	68.70 ± 6.17#	0.623 ± 0.028#	0.626 ± 0.030#	1.182 ± 0.153#
JRIP	58.16 ± 2.82#	56.78 ± 7.86#	59.93 ± 9.35#	0.579 ± 0.029#	0.582 ± 0.029#	1.265 ± 0.251#
C4.5	58.20 ± 3.66#	50.10 ± 16.48#	66.97 ± 19.25#	0.566 ± 0.042#	0.594 ± 0.031#	1.328 ± 0.522#
CART	58.18 ± 2.81#	50.48 ± 10.56#	66.33 ± 11.93#	0.575 ± 0.031#	0.592 ± 0.029#	1.327 ± 0.330#
LMT	61.53 ± 3.08#	56.76 ± 6.22#	66.65 ± 6.96#	0.614 ± 0.031#	0.664 ± 0.032#	1.198 ± 0.194#
MLP	58.32 ± 3.38#	56.57 ± 11.77#	60.29 ± 11.73#	0.578 ± 0.036#	0.624 ± 0.042#	1.266 ± 0.353#
VP	51.80 ± 5.52	59.20 ± 39.00#	48.08 ± 38.85*	0.425 ± 0.122	0.599 ± 0.048*	1.335 ± 1.169#
Bag	59.00 ± 2.75#	58.33 ± 5.35#	59.97 ± 5.49#	0.590 ± 0.028#	0.614 ± 0.036#	1.234 ± 0.162#
RF	59.71 ± 3.07#	58.40 ± 5.70#	61.30 ± 5.01#	0.597 ± 0.031#	0.628 ± 0.033#	1.219 ± 0.164#
AB	62.48 ± 2.83#	56.53 ± 4.55#	68.70 ± 6.17#	0.623 ± 0.028#	0.645 ± 0.028#	1.182 ± 0.153#

* statistically better and # statistically worse classification performance than for all variables at $P<0.05$ using Student's paired t-test

Table 10
Feature selection for 30 samples using correlation-based filter.

Variable	No. of times selected
R exp	30
EPS gr	30
InH	30
PEG	30
EPS	29
Neg	28
NCWC/R	26
ROA	11
ROE	10
EV/EBIT	5
Cash/R	3
R	2
Div/P	1
EV/EBITDA	1
no. of selected variables (mean±std.)	7.87±0.63

Table 10 summarises the results of feature selection using the combination of the correlation-based feature subset selection filter and BestFirst. The input variables are ranked according to the number of times they are selected in 30 training datasets. As can be seen from Table 10, the following variables were selected in all sample datasets as important predictors: expected growth in revenues (R exp), growth in earnings per share (EPS gr), shares held by insiders (InH), and stock price to earnings to EPS growth (PEG). It is apparent from this table that the linguistic variable Neg (frequency count of negative words) was the only linguistic variable selected (in 28 samples), meaning that in two samples no linguistic variable was used to predict fraud. Moreover, several categories of financial variables (asset structure, liquidity, and leverage ratios) were not selected at all, indicating either weak correlation with the target variable (asset structure and liquidity ratio) or strong correlations with remaining input variables, see Supplementary material for details. For example, leverage ratio BD/TA was strongly correlated with fixed assets to total assets FA/TA because fixed assets are usually financed with long-term liabilities. Similarly, liquidity was correlated with asset structure and leverage ratios because non-cash working capital (NCWC) is calculated as all non-cash current assets (total assets – fixed assets) and subtracting current liabilities. The average number of variables selected per sample was 7.87, with a standard deviation of 0.63.

Table 11 presents the performance of the nine predictive methods after feature selection. The feature selection strategy significantly improved the classification accuracies for LR, NB, SVM, MLP, RF, and AB. In addition, MC were significantly reduced for these classifiers.

To compare the classification accuracy and MC of one method against those of the other, we used Student's paired t-test, which has proved reliable for comparing different classifiers based on the mean classification performance in related studies [74]. The results of the tests in Table 12 and Table 13 indicate that the BBN and DTNB significantly outperformed other methods in terms of both classification accuracy and MC (except RF vs. BBN for MC).

As can be seen from Table 11, BBN and DTNB performed best in terms of classification accuracy, F-measure, as well as AUC. In addition, these classifiers can be considered as probabilistic white-box classifiers that represents complex relationships within the data through probability distributions [81].

In BBN, attributes have a finite set of mutually exclusive states. To each input attribute there is attached a conditional probability distribution table describing the relationship between the attribute and its parents (in this case fraudulent/non-fraudulent class; see Table 14). Thus, uncertain expert knowledge (limits further called as "red flag" and "green flag" values) can be encoded from BBN that best fits the probability distribution over the set of training data. This set of parameters that quantifies BBN was found by the K2 hill climbing algorithm with one parent. In Table 14, average probabilities are shown across the 30 datasets. The probabilities were merged into two values (> value / ≤ value) where more than two nodes were present for an attribute. For example, EPS ≤ 2.40 is a "red flag" value, indicating a high probability of fraudulent behaviour (P(EPS ≤ 2.40|class = fraud)=0.81) and a low probability of non-fraudulent behaviour (P(EPS ≤ 2.40|class = non-fraud) = 0.31). Conversely, EPS > 2.40 represents a "green flag" value, with a low probability of fraudulent (P(EPS > 2.40|class = fraud) = 0.19) and a high probability of non-fraudulent behaviour (P(EPS > 2.40|class = non-fraud) = 0.69). Notably, the probability of fraudulent behaviour sharply increases for the following values of informative variables: EPS ≤ 2.40, PEG ≤ 1.78, EV/EBIT ≤ 10.12, EV/EBITDA ≤ 7.34, EPSgr > 0.236, Rexp ≤ 0.106, NCWC/R > 0.713, Cash/R ≤ 0.79, R > 3599, InH ≤ 0.361, and Neg > 0.201. Using the probability distribution table, the overall conditional probability of financial statement fraud can be calculated. For example, P(class = fraud|EPS > 2.40, PEG ≤ 1.78, EV/EBIT ≤ 10.12, Rexp ≤ 0.106, Neg ≤ 0.201) = 0.19 × 0.81 × 0.61 × 0.89 × 0.58 × 0.5 = 0.0242, and P (class = non-fraud|EPS > 2.40, PEG ≤ 1.78, EV/EBIT ≤ 10.12, Rexp ≤ 0.106, Neg ≤ 0.201) = 0.69 × 0.32 × 0.33 × 0.40 × 0.81 × 0.5 = 0.0118, or P_{fraud} = 0.672 (0.0242/(0.0242 + 0.0118) and $P_{\text{non-fraud}}$ = 0.328 (0.0118/ (0.0242 + 0.0118) when scaled by the probability of the feature vector. Thus, the probabilities can be easily calculated in practice for the given values of the variables.

Table 11
Classification performance (mean ± std.) for 30 testing datasets using variables from feature selection.

	Accuracy	TP rate	TN rate	F-measure	AUC	MC
LR	77.31 ± 2.42*	79.11 ± 4.23*	75.68 ± 4.70	0.773 ± 0.025*	0.816 ± 0.029*	0.661 ± 0.131*
NB	61.00 ± 9.72*	52.11 ± 35.06*	70.03 ± 33.79#	0.549 ± 0.141*	0.810 ± 0.036*	1.258 ± 1.039*
BBN	90.05 ± 1.77	84.17 ± 3.08	95.91 ± 2.06	0.900 ± 0.018	0.978 ± 0.007	0.358 ± 0.082
DTNB	90.09 ± 2.13	85.53 ± 3.81	94.64 ± 2.94*	0.901 ± 0.021	0.975 ± 0.007	0.343 ± 0.106
SVM	80.50 ± 2.78*	83.25 ± 4.85*	77.98 ± 5.37*	0.805 ± 0.028*	0.806 ± 0.028*	0.555 ± 0.151*
JRIP	86.95 ± 2.29	84.85 ± 4.54	89.32 ± 7.05	0.869 ± 0.023	0.899 ± 0.024	0.410 ± 0.161
C4.5	86.60 ± 2.48	85.04 ± 5.83	88.30 ± 6.13	0.866 ± 0.025	0.911 ± 0.028*	0.416 ± 0.178
CART	87.09 ± 2.33	84.92 ± 4.40	89.33 ± 5.15*	0.871 ± 0.023	0.937 ± 0.021*	0.408 ± 0.139
LMT	86.26 ± 1.94	84.99 ± 4.18	87.62 ± 4.12	0.862 ± 0.019	0.932 ± 0.020*	0.424 ± 0.125
MLP	85.13 ± 2.54*	82.15 ± 6.37*	88.05 ± 5.86*	0.851 ± 0.025*	0.942 ± 0.016*	0.476 ± 0.186*
VP	49.59 ± 12.87	45.66 ± 29.99#	53.96 ± 29.74*	0.449 ± 0.148	0.500 ± 0.149	1.547 ± 0.897#
Bag	87.84 ± 1.33	86.54 ± 3.58	89.26 ± 3.94	0.878 ± 0.013	0.962 ± 0.007	0.377 ± 0.111
RF	88.89 ± 1.64*	87.73 ± 3.73	90.08 ± 2.92*	0.889 ± 0.017*	0.970 ± 0.006*	0.345 ± 0.104*
AB	80.50 ± 2.78*	83.25 ± 4.85*	77.98 ± 5.37	0.805 ± 0.028*	0.834 ± 0.031*	0.555 ± 0.151*

* statistically better and # statistically worse classification performance than for all variables at $P < 0.05$ using Student's paired *t*-test

Table 12
Statistical comparison (*t*-statistic) of classification accuracy.

	LR	NB	BBN	DTNB	SVM	JRIP	C4.5	CART	LMT	MLP	VP	Bag	RF
NB	−19.6#												
BBN	24.6*	43.6*											
DTNB	23.3*	42.5*	−1.7										
SVM	4.4*	23.3*	−18.9#	−17.6#									
JRIP	16.5*	34.5*	−5.2#	−3.9#	11.8*								
C4.5	17.7*	37.5*	−8.4#	−6.7#	12.2*	−1.4							
CART	16.5*	35.4*	−7.1#	−5.7#	11.5*	−1.1	0.2						
LMT	15.9*	35.2*	−8.9#	−7.4#	10.7*	−2.3#	−1.2	−1.3					
MLP	4.3*	22.7*	−18.1#	−16.9#	0.0	−11.5*	−11.8#	−11.1*	−10.3#				
VP	−26.0#	−6.8#	−49.1#	−48.0#	−29.5#	−40.2#	−43.3#	−41.2#	−41.0#	−28.8#			
Bag	18.0*	36.9*	−5.8#	−4.3#	12.9*	0.1	1.7	1.3	2.7*	12.5*	42.6*		
RF	18.8*	37.7*	−5.1#	−3.6#	13.6*	0.7	2.5*	2.0*	3.4*	13.1*	43.4*	0.7	
AB	3.6*	3.6*	−19.8#	−18.6#	−0.8	−12.6#	−13.2#	−12.4#	−11.6#	−0.8	28.7*	−13.8#	−14.5#

* statistically better and # statistically worse classification performance at $P < 0.05$ using Student's paired *t*-test

Table 13
Statistical comparison (*t*-statistic) of misclassification cost.

	LR	NB	BBN	DTNB	SVM	JRIP	C4.5	CART	LMT	MLP	VP	Bag	RF
NB	17.8#												
BBN	−20.5*	−28.4*											
DTNB	−22.5*	−29.1*	−0.3										
SVM	−4.6*	−19.9*	14.0#	15.2#									
JRIP	−18.0*	−27.1*	2.5#	3.1#	−11.7*								
C4.5	−17.8*	−27.0*	3.6#	4.4#	−11.3*	1.0							
CART	−17.0*	−26.7*	3.5#	4.1#	−10.9*	1.0	0.1						
LMT	−16.6*	−26.4*	5.0#	5.8#	−10.2*	2.3#	1.4	1.2					
MLP	−4.4*	−19.7*	13.6#	14.7#	0.0	11.4#	11.0#	10.6#	9.9#				
VP	13.1#	−9.4*	31.5#	33.6#	16.5#	29.2#	29.3#	28.4#	28.3#	16.1#			
Bag	−18.5*	−27.3*	2.8#	3.4#	−12.0*	0.1	−0.9	−0.9	−2.2*	−11.6*	−29.9*		
RF	−18.2*	−27.3*	1.8	2.3#	−12.1*	−0.6	−1.6	−1.6	−2.9*	−11.8*	−29.3*	−0.8	
AB	−3.5*	−3.5*	14.5#	15.7#	0.9	12.3#	11.9#	11.5#	10.9#	0.8	−15.3*	12.6#	12.7#

* statistically better and # statistically worse classification performance at $P < 0.05$ using Student's paired *t*-test

DTNB combines NB with induction of decision tables. The condensed form of the decision table can be used as a lookup table when making predictions. Class probability estimates are based on observed frequencies from the NB and decision tables, respectively [63]. The highly discriminative input variables were selected using search that performs a forward selection (NB) together with backward elimination (decision table). A set of decision rules to detect fraudulent/non-fraudulent firms is presented in Table 15. This is a sample from the 30 experiments for which the highest classification accuracy was achieved. Interestingly, only four input variables were selected in this sample, including not only financial variables (EPS, Rexp, and NCWC/R) but also a linguistic variable (Neg).

6. Discussion

As mentioned in the literature review, very little information was found in the extant literature on the question of combined mining financial and linguistic data for financial statement fraud prediction. The exception is the study by [30], which examined the combined effect of market value ratios and linguistic features extracted from earnings conference calls. Surprisingly, the linguistic features of earnings conference calls did not improve the performance of prediction models. On the other hand, Humpherys et al. [35] and Goel and Uzuner [36] reported a strong relationship between the linguistic features in annual reports and fraudulent behaviour. However, these studies investigated this relationship in isolation, and thus without taking financial data into considera-

Table 14
Conditional probability distribution for fraudulent/non-fraudulent firms.

Variable	Value	$P_{non\text{-}fraud}$	P_{fraud}	Value	$P_{non\text{-}fraud}$	P_{fraud}
EPS	≤ 2.40†	0.31	0.81	> 2.40	0.69	0.19
PEG	≤ 1.78†	0.32	0.81	> 1.78	0.68	0.19
EV/EBIT	≤ 10.12†	0.33	0.61	> 10.12	0.67	0.39
EV/EBITDA	≤ 7.34†	0.35	0.59	> 7.34	0.65	0.41
EPSgr	≤ 0.236	0.92	0.26	> 0.236†	0.08	0.74
Rexp	≤ 0.106†	0.40	0.89	> 0.106	0.60	0.11
ROE	≤ 0.39	0.89	0.82	> 0.39†	0.11	0.18
ROA	≤ 0.73†	0.89	0.98	> 0.73	0.11	0.02
NCWC/R	≤ 0.713	0.98	0.59	> 0.713†	0.02	0.41
Cash/R	≤ 0.79†	0.58	0.93	> 0.79	0.42	0.07
R	≤ 3599	0.84	0.41	> 3599†	0.16	0.59
Div/P	–	–	–	–	–	–
InH	≤ 0.361†	0.35	0.99	> 0.361	0.65	0.01
Neg	≤ 0.201	0.81	0.58	> 0.201†	0.19	0.42

– variable not selected as informative in BBN, † – "red flag" value

tion. The present study was therefore designed to determine the combined effect of various categories of financial data and linguistic features extracted from annual reports. To achieve this objective, we first examined the sensitivity of the performance of machine learning methods to the financial and linguistic categories of variables. Consequently, we performed feature selection to include only informative variables (with the best individual predictive ability).

Interestingly, several categories of financial variables (leverage, liquidity, and asset structure) were missing in the selected features. This is due to the fact that the filter favours features with low inter-correlations in order to decrease redundancy. Since the above-mentioned categories of financial ratios were extracted from the same financial statement, namely the balance sheet, they were strongly inter-correlated. In addition, consistent with prior literature, we observed particularly strong correlations between (1) activity ratios and liquidity and (2) reputation and asset structure. Overall, financial ratios combining information from several financial statements (income statement, balance sheet, and statement of cash flows), such as profitability and activity ratios, were more informative than those obtained from single financial statements.

The most striking result to emerge from the feature selection is that expected growth in revenues was among the most important variables. In addition, among the top seven selected features, two features are based on forecasted earnings (forward EPS and PEG ratio), and these variables are Value Line's forecasts based on limited information, both public and private. Previous empirical evidence shows that these expert forecasts have a significant advantage over traditional time series models in terms of forecast accuracy and consistency [49]. The current study found that the Value Line's forecasts of revenues and earnings are critical variables for financial statement fraud prediction. To the best of our knowledge, this result has not previously been described.

Regarding the linguistic variables, only negative sentiment in corporate annual reports was among the selected features. Although incorporating the negative sentiment, the overall tone was less informative, which can be attributed to both the strong correlation with the negative sentiment and the managerial effort to balance the negative sentiment using positive words more frequently. This is consistent with prior findings indicating that management avoids both honest reporting (which can lead to a dramatic decrease in stock price) and fraudulent reporting (which is a crime) by reducing the negative effect by use of affect modifiers like negations [82]. The present findings seem to be consistent with the study by [50], which found that negative, uncertainty, and litigious linguistic variables are all significantly linked to the 10b-5 fraud lawsuit. In contrast to these findings, however, no evidence of other linguistic variables' effects was detected. A reasonable explanation for this result may be that Loughran and McDonald [50] used data from a different period (1994–2004) and, therefore, the results are likely to be affected by substantial changes in financial reporting standards. The effect of these changes has been demonstrated for bankruptcy prediction models [83].

When interpreting the results of machine learning methods, several issues arise. The most interesting finding was that ensemble methods outperformed the remaining methods in terms of *TP* rate, with the highest rate of fraudulent firms correctly classified as fraudulent. This performance metric has been proved to be crucial for fraud prediction models due to the significantly higher MC associated with fraud. Achieving a high prediction accuracy at detecting fraudulent firms has therefore become a central task in intelligent financial statement fraud prediction. On the other hand, BBN and DTNB provided the highest accuracy on non-fraudulent firms

Table 15
Decision table to identify non-fraudulent (0) and fraudulent (1) firms.

EPS	Rexp	NCWC/R	Neg	Decision	EPS	Rexp	NCWC/R	Neg	Decision
2.4–2.44	0.106–0.109	> 0.97	> 0.018	0	≤ 2.35	≤ 0.096	≤ 0.71	≤ 0.018	0
> 2.44	> 0.109	> 0.97	≤ 0.018	0	> 2.44	≤ 0.096	≤ 0.71	≤ 0.018	0
2.35–2.4	0.099–0.106	0.71–0.97	> 0.018	0	2.4–2.44	≤ 0.096	≤ 0.71	≤ 0.018	0
2.35–2.4	> 0.109	≤ 0.71	> 0.018	0	> 2.44	> 0.109	> 0.97	> 0.018	1
> 2.44	≤ 0.096	> 0.97	> 0.018	0	2.35–2.4	> 0.109	0.71–0.97	> 0.018	1
2.35–2.4	≤ 0.096	> 0.97	> 0.018	0	≤ 2.35	> 0.109	≤ 0.71	> 0.018	1
2.4–2.44	> 0.109	≤ 0.71	> 0.018	0	≤ 2.35	0.096–0.099	0.71–0.97	> 0.018	1
> 2.44	> 0.109	≤ 0.71	> 0.018	0	> 2.44	0.096–0.099	0.71–0.97	> 0.018	1
≤ 2.35	0.106–0.109	> 0.97	≤ 0.018	0	2.35–2.4	> 0.109	0.71–0.97	≤ 0.018	1
> 2.44	0.106–0.109	≤ 0.71	> 0.018	0	2.35–2.4	0.096–0.099	0.71–0.97	> 0.018	1
≤ 2.35	0.106–0.109	≤ 0.71	> 0.018	0	2.35–2.4	≤ 0.096	0.71–0.97	> 0.018	1
2.4–2.44	0.106–0.109	≤ 0.71	> 0.018	0	≤ 2.35	0.096–0.099	≤ 0.71	> 0.018	1
≤ 2.35	0.099–0.106	≤ 0.71	> 0.018	0	> 2.44	0.096–0.099	≤ 0.71	> 0.018	1
2.4–2.44	> 0.109	≤ 0.71	≤ 0.018	0	2.35–2.4	0.096–0.099	≤ 0.71	> 0.018	1
> 2.44	> 0.109	≤ 0.71	≤ 0.018	0	≤ 2.35	0.096–0.099	0.71–0.97	≤ 0.018	1
≤ 2.35	> 0.109	≤ 0.71	≤ 0.018	0	> 2.44	≤ 0.096	≤ 0.71	> 0.018	1
> 2.44	0.106–0.109	≤ 0.71	≤ 0.018	0	2.35–2.4	0.096–0.099	0.71–0.97	≤ 0.018	1
2.35–2.4	≤ 0.096	≤ 0.71	> 0.018	0	≤ 2.35	≤ 0.096	≤ 0.71	> 0.018	1
≤ 2.35	0.106–0.109	≤ 0.71	≤ 0.018	0	2.35–2.4	≤ 0.096	0.71–0.97	≤ 0.018	1
> 2.44	0.096–0.099	0.71–0.97	≤ 0.018	0	> 2.44	0.096–0.099	≤ 0.71	≤ 0.018	1
2.4–2.44	≤ 0.096	≤ 0.71	> 0.018	0	2.35–2.4	0.096–0.099	≤ 0.71	≤ 0.018	1
2.4–2.44	0.106–0.109	≤ 0.71	≤ 0.018	0	≤ 2.35	0.096–0.099	≤ 0.71	≤ 0.018	1
≤ 2.35	0.099–0.106	≤ 0.71	≤ 0.018	0	> 2.44	0.096–0.099	≤ 0.71	≤ 0.018	1
2.35–2.4	≤ 0.096	≤ 0.71	≤ 0.018	0					

(*TN* rate). This suggests that correct prediction of non-fraudulent firms is less complex, whereas the detection of fraudulent firms requires more complex (and less interpretable) machine learning methods. As a result, we were not able to develop accurate "red flag" values of predictive indicators suggesting the presence of fraud. In contrast, to the best of our knowledge this is the first study developing "green flag" values, for which fraud is likely absent. Any violation of these critical values should be a reason for further investigation by an auditor. Thus, the "green flag" values can possibly support the decisions of auditors during client selection or audit planning. In addition, the decision rules generated by the DTNB can be used to detect fraudulent/non-fraudulent firms with the mean *TP* rate above 85 % (and *TN* rate around 95%).

The superior performance of both the BBN / DTNB and the ensembles of decision trees over SVM and MLP can be explained by a relatively low dimensionality achieved using feature selection, preventing overfitting of the training data. In addition, the selection of non-correlated features may also favour machine learning methods with conditional independence assumption (like BBN and DTNB). Furthermore, a relatively large dataset (compared with previous related studies) also favours machine learning methods that minimise empirical risk. Finally, another possible explanation for the good performance of the ensembles of decision trees is the frequent presence of outliers in the data.

The results of the sensitivity analysis indicate that corporate reputation and market value ratios significantly improve the performance of all machine learning methods. In contrast, profitability ratios and linguistic variables were important for the prediction performance of BBN, suggesting that these variables are crucial for the prediction of non-fraudulent firms in particular. This indicates that both financial statements and text in annual reports can be effectively utilised to detect non-fraudulent firms. However, non-annual report data are necessary when attempting to locate and identify fraudulent firms. This finding has important implications for selecting variables when developing intelligent financial statement fraud detection systems.

The results of this study could also be a useful aid for government and regulatory bodies when developing targeted interventions aimed at firms that are more susceptible to financial statement fraud. Such decision aids are also important for external auditors, most of whom have limited experience in detecting fraud and inadequate auditing resources [35]. Unnecessary audit costs are positively correlated with the *FP* rate. Failing to detect fraud (*FN* rate), on the other hand, incurs significant financial losses. In this study, we adopted the cost ratio of 1:2, but this ratio largely depends on the type of user. For instance, Abbasi et al. [5] recommend different ratios for investors (1:20) and regulators (1:10). Whereas the costs for regulators are similar to those for auditors, investors suffer (1) financial losses due to decrease in stock value when the fraud is discovered or (2) opportunity cost when failing to invest in a non-fraudulent firm. A higher cost ratio for investors can also be explained by a portfolio approach that minimises opportunity cost [30]. Unnecessary costly audit procedures, on the other hand, decrease the ratio in case of external auditors. In future investigations, it might therefore be possible to use different performance metrics specifically developed for various user categories.

7. Conclusion

This study examined whether an improved financial fraud detection tool could be developed by combining specific features derived from financial information and corporate annual reports. Our work has led us to conclude that a low relative frequency of negative words in annual reports may indicate non-fraudulent firms. However, to detect fraudulent firms, it is necessary to use information extracted from both publicly available financial statements and analysts' forecasts of revenues and earnings.

The current research was not specifically designed to evaluate the textual content of corporate annual reports, and thus we omitted several text mining analyses such as bag-of-words [84], part-of-speech tagging [85] and concept extraction [86]. Further experimental investigations are therefore needed to estimate their effect in combination with the determinants examined in the present study. Moreover, further experimentation with information asymmetries arising from industry-level complexities is also recommended [87].

To select only informative features, we performed feature selection using a filter method. One obvious advantage of this approach is that filters operate independently of any learning algorithm. However, it would be interesting to assess the effects of various feature selection methods on classification performance. Furthermore, the present study was limited by the use of a balanced sample of fraudulent and non-fraudulent firms. Future studies could assess the effect of increasing the number of non-fraudulent firms on training data. The use of machine learning methods designed specifically for imbalanced data are therefore strongly recommended. Finally, our results are encouraging and should be validated by a larger sample size in further research to increase the significance of the results. For longer time series, it would also be interesting to investigate the effect of concept drift recently observed in financial distress prediction [74].

Acknowledgments

We gratefully acknowledge the help provided by constructive comments of the anonymous referees. This work was supported by the scientific research project of the Czech Sciences Foundation Grant No: 16-19590S.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.knosys.2017.05.001.

References

[1] K. Nguyen, Financial Statement Fraud: Motives, Methods, Cases and Detection, Universal-Publishers, Boca Raton, 2010.
[2] S.Y. Huang, R.H. Tsaih, F. Yu, Topological pattern discovery and feature extraction for fraudulent financial reporting, Expert Syst. Appl. 41 (2014) 4360–4372, doi:10.1016/j.eswa.2014.01.012.
[3] G.K. Agarwal, Y. Medury, Internal auditor as accounting fraud buster, IUP J. Account. Res. Audit Pract. 13 (2014) 7–29.
[4] M.S. Beasley, J.V Carcello, D.R. Hermanson, T.L. Neal, Fraudulent Financial Reporting, Committee of Sponsoring Organizations of the Treadway Commission, Jersey City, 2010.
[5] A. Abbasi, C. Albrecht, A. Vance, J. Hansen, Metafraud: a meta-learning framework for detecting financial fraud, MIS Q. 36 (2012) 1293–1327.
[6] M. Beneish, The detection of earnings manipulation, Financ. Anal. J. 5 (1999) 24–36.
[7] A. Dyck, A. Morse, L. Zingales, Who blows the whistle on corporate fraud? J. Finance. 65 (2010) 2213–2253, doi:10.1111/j.1540-6261.2010.01614.x.
[8] D. Moore, P. Tetlock, L. Tanlu, Conflicts of interest and the case of auditor independence: moral seduction and strategic issue cycling, Acad. Manag. Rev. 31 (2006) 10–29.
[9] J. West, M. Bhattacharya, Intelligent financial fraud detection: a comprehensive review, Comput. Secur. 57 (2016) 47–66, doi:10.1016/j.cose.2015.09.005.
[10] W.S. Albrecht, C. Albrecht, C.C. Albrecht, Current trends in fraud and its detection, Inf. Secur. J. A Glob. Perspect. 17 (2008) 2–12, doi:10.1080/19393550801934331.
[11] E. Kirkos, C. Spathis, Y. Manolopoulos, Data mining techniques for the detection of fraudulent financial statements, Expert Syst. Appl. 32 (2007) 995–1003, doi:10.1016/j.eswa.2006.02.016.
[12] C.C. Lin, A.A. Chiu, S.Y. Huang, D.C. Yen, Detecting the financial statement fraud: the analysis of the differences between data mining techniques and experts' judgments, Knowl. Based Syst. 89 (2015) 459–470, doi:10.1016/j.knosys.2015.08.011.
[13] K. Fanning, K.O. Cogger, Neural network detection of management fraud using published financial data, Int. J. Intell. Syst. Account. Financ. Manag. 7 (1998) 21–41, doi:10.1002/j.1099-1174.1995.tb00084.x.

[14] K. Fanning, K. Cogger, R. Srivastava, Detection of management fraud: a neural network approach, Int. J. Intell. Syst. Account. Financ. Manag. 4 (1995) 113–126, doi:10.1002/j.1099-1174.1995.tb00084.x.
[15] S. Kotsiantis, E. Koumanakos, D. Tzelepis, Forecasting fraudulent financial statements using data mining, Int. J. Comput. Intell. 3 (2006) 104–110.
[16] P.F. Pai, M.F. Hsu, M.C. Wang, A support vector machine-based model for detecting top management fraud, Knowl. Based Syst 24 (2011) 314–321, doi:10.1016/j.knosys.2010.10.003.
[17] S. Huang, Fraud detection model by using support vector machine techniques, Int. J. Digit. Content Technol. Its Appl. 7 (2013) 32–42.
[18] B. Hoogs, T. Kiehl, C. Lacomb, D. Senturk, A genetic algorithm approach to detecting temporal patterns indicative of financial statement fraud, Intell. Syst. Account. Financ. Manag. 15 (2007) 41–56, doi:10.1002/isaf.
[19] G. Pradeep, V. Ravi, K. Nandan, B. Deekshatulu, Fraud detection in financial statements using evolutionary computation based rule miners, in: B.K. Panigrahi, P.N. Suganthan, S. Das (Eds.), 5th Int. Conf. Swarm, Evol. Memetic Comput., Springer, Bhubaneswar, 2014, pp. 239–250, doi:10.1007/978-3-319-20294-5_21.
[20] M. Cecchini, H. Aytug, G.J. Koehler, P. Pathak, Making words work: using financial text as a predictor of financial events, Decis. Support Syst. 50 (2010) 164–175, doi:10.1016/j.dss.2010.07.012.
[21] S. Goel, J. Gangolly, S.R. Faerman, O. Uzuner, Can linguistic predictors detect fraudulent financial filings? J. Emerg. Technol. Account. 7 (2010) 25–46, doi:10.2308/jeta.2010.7.1.25.
[22] L. Purda, D. Skillicorn, Accounting variables, deception, and a bag of words: assessing the tools of fraud detection, Contemp. Account. Res. 32 (2015) 1193–1223, doi:10.1111/1911-3846.12089.
[23] S. Bhattacharyya, S. Jha, K. Tharakunnel, J.C. Westland, Data mining for credit card fraud: a comparative study, Decis. Support Syst. 50 (2011) 602–613, doi:10.1016/j.dss.2010.08.008.
[24] D. Olszewski, Fraud detection using self-organizing map visualizing the user profiles, Knowl. Based Syst. 70 (2014) 324–334, doi:10.1016/j.knosys.2014.07.008.
[25] M. Jans, J.M. Van Der Werf, N. Lybaert, K. Vanhoof, A business process mining application for internal transaction fraud mitigation, Expert Syst. Appl. 38 (2011) 13351–13359, doi:10.1016/j.eswa.2011.04.159.
[26] L. Bermúdez, J.M. Pérez, M. Ayuso, E. Gómez, F.J. Vázquez, A Bayesian dichotomous model with asymmetric link for fraud in insurance, Insur. Math. Econ. 42 (2008) 779–786, doi:10.1016/j.insmatheco.2007.08.002.
[27] E.W.T. Ngai, Y. Hu, Y.H. Wong, Y. Chen, X. Sun, The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature, Decis. Support Syst. 50 (2011) 559–569, doi:10.1016/j.dss.2010.08.006.
[28] B.S. Kumar, V. Ravi, A survey of the applications of text mining in financial domain, Knowl. Based Syst. 114 (2016) 128–147, doi:10.1016/j.knosys.2016.10.003.
[29] G.L. Gray, R.S. Debreceny, A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits, Int. J. Account. Inf. Syst. 15 (2014) 357–380, doi:10.1016/j.accinf.2014.05.006.
[30] C.S. Throckmorton, W.J. Mayew, M. Venkatachalam, L.M. Collins, Financial fraud detection using vocal, linguistic and financial cues, Decis. Support Syst. 74 (2015) 78–87, doi:10.1016/j.dss.2015.04.006.
[31] J. Sun, H. Li, Q.H. Huang, K.Y. He, Predicting financial distress and corporate failure: a review from the state-of-the-art definitions, modeling, sampling, and featuring approaches, Knowl. Based Syst 57 (2014) 41–56, doi:10.1016/j.knosys.2013.12.006.
[32] P. Ravisankar, V. Ravi, G. Raghava Rao, I. Bose, Detection of financial statement fraud and feature selection using data mining techniques, Decis. Support Syst. 50 (2011) 491–500, doi:10.1016/j.dss.2010.11.006.
[33] C. Gaganis, Classification techniques for the identification of falsified financial statements: a comparative analysis, Intell. Syst. Account. Financ. Manag. 16 (2009) 207–229, doi:10.1002/isaf.303.
[34] P.M. Dechow, W. Ge, C.R. Larson, R.G. Sloan, Predicting material accounting misstatements, Contemp. Account. Res. 28 (2011) 17–82, doi:10.1111/j.1911-3846.2010.01041.x.
[35] S.L. Humpherys, K.C. Moffitt, M.B. Burns, J.K. Burgoon, W.F. Felix, Identification of fraudulent financial statements using linguistic credibility analysis, Decis. Support Syst. 50 (2011) 585–594, doi:10.1016/j.dss.2010.08.009.
[36] S. Goel, O. Uzuner, Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports, Intell. Syst. Account. Financ. (2016), doi:10.1002/isaf.1392.
[37] S. Minhas, A. Hussain, From spin to swindle: Identifying falsification in financial text, Cognit. Comput. 8 (2016) 729–745, doi:10.1007/s12559-016-9413-9.
[38] J.L. Hobson, W.J. Mayew, M. Venkatachalam, Analyzing speech to detect financial misreporting, J. Account. Res. 50 (2012) 349–392, doi:10.1111/j.1475-679X.2011.00433.x.
[39] E.F. Zainudin, H.A. Hashim, Detecting fraudulent financial reporting using financial ratio, J. Financ. Report. Account. 14 (2016) 266–278, doi:10.1108/JFRA-05-2015-0053.
[40] S. Chen, Y.J.J. Goo, Z. De Shen, A hybrid approach of stepwise regression, logistic regression, support vector machine, and decision tree for forecasting fraudulent financial statements, Sci. World J. (2014) 1–9, doi:10.1155/2014/968712.
[41] J. Perols, Financial statement fraud detection: an analysis of statistical and machine learning algorithms, Audit. A J. Pract. Theory. 30 (2011) 19–50, doi:10.2308/ajpt-50009.
[42] Y.J. Kim, B. Baik, S. Cho, Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning, Expert Syst. Appl. 62 (2016) 32–43, doi:10.1016/j.eswa.2016.06.016.
[43] X. Li, W. Xu, X. Tian, How to protect investors? A GA-based DWD approach for financial statement fraud detection, in: 2014 IEEE Int. Conf. Syst. Man, Cybern., San Diego, IEEE, 2014, pp. 3548–3554, doi:10.1109/SMC.2014.6974480.
[44] J.M. Karpoff, A. Koester, D.S. Lee, G.S. Martin, Database challenges in financial misconduct research, Work. Pap. (2014) 1–66 http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2112569.
[45] O.S. Persons, Using financial statement data to identify factors associated with fraudulent financial reporting, J. Appl. Bus. Res. 11 (1995) 3846.
[46] B. Bai, J. Yen, X. Yang, False financial statements: characteristics of china's listed companies and cart detecting approach, Int. J. Inf. Technol. Decis. Mak. 7 (2008) 339–359, doi:10.1142/S0219622008002958.
[47] P. Hajek, K. Michalak, Feature selection in corporate credit rating prediction, Knowl. Based Syst. 51 (2013) 72–84, doi:10.1016/j.knosys.2013.07.008.
[48] P.K. Sen, Ownership incentives and management fraud, J. Bus. Financ. Account. 34 (2007) 1123–1140, doi:10.1111/j.1468-5957.2007.02026.x.
[49] Y. Zhang, B. Alexander, Half a century of research on value line: a comprehensive review, Manag. Financ. 42 (2016) 799–816. http://dx.doi.org/10.1108/MF-02-2015-0053.
[50] T. Loughran, B. McDonald, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, J. Finance. 66 (2011) 35–65, doi:10.1111/j.1540-6261.2010.01625.x.
[51] M.L. Newman, J.W. Pennebaker, D.S. Berry, J.M. Richards, Lying words: predicting deception from linguistic cues, Personal. Soc. Psychol. Bull. 29 (2003) 665–675, doi:10.1177/0146167203251529.
[52] S.H. Adams, J.P. Jarvis, Indicators of veracity and deception: an analysis of written statements made to police, Int. J. Speech Lang. Law. 13 (2006), doi:10.1558/sll.2006.13.1.1.
[53] A. Vrij, Detecting Lies and Deceit: Pitfalls and Opportunities, 2nd ed., John Wiley & Sons, Chichester, 2008.
[54] D.F. Larcker, A.A. Zakolyukina, Detecting deceptive discussions in conference calls, J. Account. Res. 50 (2012) 495–540, doi:10.1111/j.1475-679X.2012.00450.x.
[55] A. Bodnaruk, T. Loughran, B. McDonald, Using 10-K text to gauge financial constraints, J. Financ. Quant. Anal. 50 (2015) 623–646. http://dx.doi.org/10.2139/ssrn.2331544.
[56] E. Henry, Are investors influenced by how earnings press releases are written? J. Bus. Commun. 45 (2008) 363–407, doi:10.1177/0021943608319388.
[57] I. Guyon, A. Elisseeff, An introduction to feature extraction, in: I. Guyon, M. Nikravesh, S. Gunn, L.A. Zadeh (Eds.), Featur. Extr. - Found. Appl., Springer, Berlin Heidelberg, 2006, pp. 1–25, doi:10.1007/978-3-540-35488-8.
[58] M.A. Hall, Correlation-Based Feature Selection For Machine Learning, The University of Waikato, 1999.
[59] R. Dechter, J. Pearl, Generalized best-first search strategies and the optimality of A*, J. ACM. 32 (1985) 505–536, doi:10.1145/3828.3830.
[60] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, 2nd ed., John Wiley & Sons, New York, 2000, doi:10.2307/2074954.
[61] G. John, P. Langley, Estimating continuous distributions in Bayesian classifiers, in: Elev. Conf. Uncertain. Artif. Intell., San Francisco, Morgan Kaufmann, 1995, pp. 338–345.
[62] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, Waltham, 2011.
[63] M. Hall, E. Frank, Combining naive Bayes and decision tables, Intelligence (2008) 2–3.
[64] J.C. Platt, Fast training of support vector machines using sequential minimal optimization, Adv. Kernel Methods (1998) 185–208, doi:10.1109/ISKE.2008.4731075.
[65] J. Quinlan, C4.5: Programs For Machine Learning, Morgan Kaufmann, San Mateo, 1992.
[66] D. Rumelhart, G. Hinton, R. Williams, Learning representations by back-propagating errors, in: T.A. Polk, C.M. Seifert (Eds.), Cogn. Model., MIT Press, Cambridge, 1988, pp. 213–222.
[67] Y. Freund, R.E. Schapire, Large margin classification using the perceptron algorithm, Mach. Learn. 37 (1999) 277–296, doi:10.1023/A:1007662407062.
[68] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140, doi:10.1007/BF00058655.
[69] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, doi:10.1023/A:1010933404324.
[70] Y. Freund, R.R.E. Schapire, Experiments with a new boosting algorithm, Int. Conf. Mach. Learn. (1996) 148–156 10.1.1.133.1040.
[71] D. Heckerman, A tutorial on learning with Bayesian networks, Innov. Bayesian Networks, 2008.
[72] D.E. Rumelhart, J.L. McClelland, R.J. Williams, Parallel Distributed Processing: Explorations In The Microstructure Of Cognition, MIT Press, Cambridge, 1986.
[73] C. Liu, Y. Chan, S. Hasnain, A. Kazmi, H. Fu, Financial fraud detection model: based on random forest, Int. J. Econ. Financ. 7 (2015) 178–188, doi:10.5539/ijef.v7n7p178.Financial.
[74] J. Sun, H. Fujita, P. Chen, H. Li, Dynamic financial distress prediction with concept drift based on time weighting combined with adaboost support vector machine ensemble, Knowl. Based Syst. 120 (2017) 4–14, doi:10.1016/j.knosys.2016.12.019.
[75] J. West, M. Bhattacharya, Mining financial statement fraud: an analysis of some experimental issues, in: IEEE 10th Conf. Ind. Electron. Appl., IEEE, Auckland, 2015, pp. 461–466, doi:10.1109/ICIEA.2015.7334157.

[76] J. Hansen, J. McDonald, W. Messier, T. Bell, A generalized qualitative-response model and the analysis of management fraud, Manage. Sci. 42 (1996) 1022–1032, doi:10.1287/mnsc.42.7.1022.
[77] J. Bauer, V. Agarwal, Are hazard models superior to traditional bankruptcy prediction approaches? A comprehensive test, J. Bank. Financ. 40 (2014) 432–442, doi:10.1016/j.jbankfin.2013.12.013.
[78] Association of Certified Faud Examiners, Report to the nations on occupational fraud and abuse, Glob. Fraud Study 9 (2016) 1–91. http://dx.doi.org/10.2139/ssrn.2222608.
[79] Financial Executives International, Audit Fee Survey Report, Financial Executives Research Foundation, Morristown, 2016.
[80] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (2006) 861–874, doi:10.1016/j.patrec.2005.10.010.
[81] K. Dejaeger, T. Verbraken, B. Baesens, Toward comprehensible software fault prediction models using Bayesian network classifiers, IEEE Trans. Softw. Eng. 39 (2013) 237–257, doi:10.1109/TSE.2012.20.
[82] F.H. Glancy, S.B. Yadav, A computational model for financial reporting fraud detection, Decis. Support Syst. 50 (2011) 595–601, doi:10.1016/j.dss.2010.08.010.
[83] W.H. Beaver, M. Correia, M.F. McNichols, Do differences in financial reporting attributes impair the predictive ability of financial ratios for bankruptcy? Rev. Account. Stud. 17 (2012) 969–1010, doi:10.1007/s11142-012-9186-7.
[84] P. Saha, I. Bose, A. Mahanti, A knowledge based scheme for risk assessment in loan processing by banks, Decis. Support Syst. 84 (2016) 78–88, doi:10.1016/j.dss.2016.02.002.
[85] R.P. Schumaker, Y. Zhang, C.N. Huang, H. Chen, Evaluating sentiment in financial news articles, Decis. Support Syst. 53 (2012) 458–464, doi:10.1016/j.dss.2012.03.001.
[86] S. Feuerriegel, A. Ratku, Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation, in: T.X. Bui, R.H. Sprague (Eds.), 49th Hawaii Int. Conf. Syst. Sci., IEEE, Kauai, 2016, pp. 1072–1081, doi:10.1109/HICSS.2016.137.
[87] H.A. Ndofor, C. Wesley, R.L. Priem, Providing CEOs with opportunities to cheat: The effects of complexity-based information asymmetries on financial reporting fraud, J. Manage. 41 (2015) 1774–1797, doi:10.1177/0149206312471395.
[88] W. Dong, S. Liao, B. Fang, X. Cheng, C. Zhu, W. Fan, The detection of fraudulent financial statements: an integrated language model approach, in: Pacific Asia Conf. Inf. Syst. PACIS 2014, 2014, p. 383.
[89] X.-P. Song, Z.-H. Hu, J.-G. Du, Z.-H. Sheng, Application of machine learning methods to risk assessment of financial statement fraud: evidence from China, J. Forecast. 33 (2014) 611–626, doi:10.1002/for.2294.
[90] F. Chen, D. Chi, J. Zhu, Application of random forest, rough Set theory, decision tree and neural network to detect financial statement fraud – taking corporate governance into consideration, in: D.-S. Huang, V. Bevilacqua, P. Premaratne (Eds.), 10th Int. Conf. Intell. Comput. Theory, ICIC 2014, Springer, Taiyuan, 2014, pp. 221–234, doi:10.1007/978-3-319-09333-8_24.