International Journal of Accounting Information Systems 26 (2017) 32–45

Contents lists available at ScienceDirect

International Journal of Accounting Information Systems

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

Enhancement of fraud detection for narratives in annual reports

Yuh-Jen Chen[a], Chun-Han Wu[b], Yuh-Min Chen[b], Hsin-Ying Li[a], Huei-Kuen Chen[c,*]

[a] Department of Accounting and Information Systems, National Kaohsiung First University of Science and Technology, Kaohsiung, Taiwan, ROC
[b] Institute of Manufacturing Information and Systems, National Cheng Kung University, Tainan, Taiwan, ROC
[c] College of Finance and Banking, National Kaohsiung First University of Science and Technology, Kaohsiung, Taiwan, ROC

ARTICLE INFO

Keywords:
Narratives
Annual reports
Fraud detection
Natural language processing (NLP)
Queen genetic algorithm (QGA)
Support vector machine (SVM)

ABSTRACT

Annual reports present the activities of a listed company in terms of its operational performance, financial conditions, and social responsibilities. These reports are a valuable reference for numerous investors, creditors, and other accounting information end users. However, many annual reports exaggerate enterprise activities to raise investors' capital and support from financial institutions, thereby diminishing the usefulness of such reports. Effectively detecting fraud in the annual report of a company is thus a priority concern during an audit.

Therefore, this work integrates natural language processing (NLP), queen genetic algorithm (QGA) and support vector machine (SVM) to develop a fraud detection method for narratives in annual reports, such as reports to shareholders, and thereby enhance the fraud detection accuracy and reduce investors' investment risks. To achieve the above-mentioned objective, a process of fraud detection for narratives in annual reports is first designed. Techniques related to fraud detection for the narratives in annual reports are then developed. Finally, the proposed fraud detection method is demonstrated and evaluated.

1. Introduction

In addition to orienting investors the operational performance, risks, and growth potential of an enterprise, an annual report provides information to creditors and suppliers of the debt payment capability of an enterprise and facilitates governmental auditing of company revenues for tax purposes. An annual report also allows an enterprise to reduce information asymmetry with end users such as investors. However, some annual reports might exaggerate enterprise activities to raise investor capital and support from financial institutions, thereby diminishing the usefulness of such reports. Effectively detecting fraud in the annual report of a company is thus a priority for auditors, investors, and creditors.

The studies of fraud detection for financial statements can be classified into two categories. One category is to develop some detection methods to detect potentially fraudulent financial reports (Kaminski et al., 2004; Zhou et al., 2004), including numerical and textual financial reports. The other is to focus on identifying potential fraudulent features, such as backdating, which can be used for efficiently detecting fraudulent financial statements (Hake, 2005; Siegel, 2007; Tillman and Indergaard, 2003).

Beattie et al. (2004) indicated that the narratives in annual reports comprised eight main topics (financial data, operating data, management analysis, forward-looking information, information about management and shareholders, objectives and strategy, description of business and industry structure). Yekini et al. (2016) stated that, in the UK, the Companies Act 2006 and the amendments to this Act introduced in 2013 required large and medium listed companies to incorporate certain sections in their annual reports. These included the strategic report/business review section (covering business description, issues related to performance, principal

[*] Corresponding author.
E-mail address: u0447910@nkfust.edu.tw (H.-K. Chen).

http://dx.doi.org/10.1016/j.accinf.2017.06.004
Received 20 April 2016; Received in revised form 22 May 2017; Accepted 19 June 2017
Available online 19 July 2017
1467-0895/

risks, position, trends and factors, and key performance indicators), the corporate social responsibility statement (describing environmental, employee and community issues), the directors' reports, the directors' remuneration reports, and the statement of directors' responsibilities. Wisniewski and Yekini (2015) mentioned that the Companies Act (2006) mandated large and medium quoted companies to include a business review section covering a description of company business, its performance, principal risks, position, trends and factors, as well as financial and non-financial key performance indicators (KPIs). These narratives provide a rich set of data that are used by investors and creditors to evaluate the risk associated with companies. However, companies could use these narratives to potentially fraudulently mislead investors and creditors.

Various fraud detection methods for numerical and textual financial reports/statements have been recently developed. For fraud detection in numerical financial reports/statements, Kirkos et al. (2007) used data mining classification techniques to efficiently detect firms' fraudulent financial statements and identify several factors related to fraudulent financial statements. Huang et al. (2008) developed a mechanism for innovative fraud detection based on Zipf's Law to assist auditors in examining the vast volumes of operational datasets and identifying possibly fraudulent records. Ravisankar et al. (2011) used various techniques of data mining, including multilayer feed forward neural network, support vector machines, genetic programming, group method of data handling, logistic regression, and probabilistic neural network, to detect fraudulent financial statements of companies. Dechow et al. (2011) established a comprehensive database of financial misstatements and provided it for researchers to promote research on earnings misstatements. Moreover, the logistic model for predicting misstatements was developed through analyzing the financial features of misstating firms. Gupta and Gill (2012) proposed a data mining framework to prevent and detect financial statement fraud. In the framework, data mining techniques were employed to use past fraudulent cases to establish prevention and detection models for fraud risks and financial statement fraud. Alden et al. (2012) adopted a genetic algorithm and a modern estimation of distribution algorithm to develop the fuzzy rule-based classifiers for detecting financial statements. In the demonstration, the two algorithms had a better ability to identify fraudulent financial statements than those of a traditional logistic regression model.

In addition to focusing on the financial information contained in the annual reports, Brazel et al. (2009) investigated how auditors could effectively utilize nonfinancial indicators for measuring the reasonableness of financial performance for financial statement fraud detection. Debreceny and Gray (2010) explored the applications of data mining techniques to effectively and efficiently detect fraud in journal entries. Pai et al. (2011) combined sequential forward selection, support vector machine, and a classification and regression tree to devise a support vector machine-based fraud warning model to decrease the related risks caused by inexperienced auditors who were in detecting fraud for financial statements.

Fraud detection in textual financial reports/statements was examined by Glancy and Yadav (2011) who developed a computational fraud detection model, in which a quantitative approach on textual data was used for detecting fraud in financial reports. Humpherys et al. (2011) proposed a novel approach, which applied text mining methods to identify fraud in the Management's Discussion and Analysis of the Form10-K to assist auditors in measuring the fraud risk.

Additionally, studies on content analysis of annual reports or accounting information as well as fraud detection through narrative disclosures or linguistics also have been developed. For example, Edward (1984) used annual report content analysis to explore corporate strategy and factors in risk and return. In the experimental report, three industries of food processing, computer peripherals, and containers were given to demonstrate that a negative correlation of risk and return between companies in industries. Breton and Taffler (2001)) explored the importance of accounting measures, compared with non-financial information utilized by stock analysts in recommending stocks through analyzing companies' report contents. The authors concluded that accounting information was the most important information item for analysts. Zhou et al. (2007)) developed a system of automated linguistics based cues for deception detection. In the experiment, the automated linguistics based cues in the context of text-based asynchronous computer mediated communication were demonstrated to be effective in the detection of deception. Churyk et al. (2009)) applied the content analysis to the management discussion and analysis in the annual report to identify potential indicators of fraud for early detection of fraud. The findings indicated that qualitative methods of deception detection could provide a useful method for detecting fraud.

Tausczik and Pennebaker (2010)) examined various computerized text analysis methods and explained how linguistic inquiry and word count (LIWC) were created and validated. The experimental results indicated that the LIWC had the ability to detect signification in attentional focus, emotionality, social relationships, thinking styles, and individual differences. Li et al. (2012)) used LIWC to compare the linguistic and psychological term uses in English and Chinese languages. In the experiment, the technique of principal component analysis was employed and five linguistic and psychological components were identified. Lee et al. (2013)) described a process of model building and validation for early fraud prediction according to the narrative disclosures in annual reports. They used content analysis to examine the management discussion and analysis in the annual reports to identify important qualitative fraud risk factors.

For detecting narrative fraud in annual reports, many recent studies proposed various text mining techniques to enhance the detection accuracy. The average accuracy of these studies on detecting narrative fraud in annual reports was about 72%, as shown in Table 1. Moreover, the LIWC has been proven to be a psychology tool that is increasingly being used for content analysis (Pennebaker et al., 2007; Pennebaker et al., 2001). Several studies with LIWC-based text analysis methods were proposed to count the frequency of occurrence of words in psychology, such as emotional words being used for calculating the percentage of relative use. In the LIWC program, the dictionaries were the core feature. When the dictionaries were first established, emotion words in a text were only considered and computed by the computer. For other psychological word categories, human judgement was required for evaluating which words were best suited for these categories. This situation not only increased the cost of human judgement for creating various psychological words, but also did not allow other psychological word categories to be automatically created and updated for the establishment and growth of domain dictionary.

Table 1
Existing studies on fraud detection/prediction for financial information.

	Numerical financial data	Textual financial data	Detection/prediction model	Detection/prediction accuracy
Kirkos et al., 2007	✓		Bayesian belief Networks	90.30%
Huang et al. (2008)	✓		Zipf's Law	96.45%
Skousen and Wright (2008) Claude (1948)	✓		Logit regression	69.77%
Glancy and Yadav (2011)		✓	Text mining	83.87%
Humpherys et al. (2011)		✓	Text mining	67.30%
Pai et al. (2011)	✓		Classification and regression tree	92.00%
Pennebaker et al. (2001)	✓		Genetic programming	92.68%
Alden et al. (2012)	✓		Evolutionary algorithm	64.46%
Dechow et al. (2011)	✓		Logistic regression	63.00%
Lee et al. (2013)		✓	Content analysis-based stepwise model	64.80%
This Study Chen et al. (2017)		✓	NLP, QGA, and SVM	85.25%

This work adopts another natural language processing (NLP) program and integrates queen genetic algorithm (QGA) and support vector machine (SVM) to develop a fraud detection method for narratives in annual reports. This method overcomes the limitation of the need to manually create psychological word categories and can help investors detect fraudulent narratives in annual reports and reduce investment risks. To achieve the above-mentioned objective, a process of fraud detection for narratives in annual reports is first designed. Next, techniques related to fraud detection for narratives in annual reports are developed. Finally, the proposed fraud detection method is demonstrated and evaluated. Fraud detection-related techniques for narratives in annual reports consist mainly of establishing a fraudulent feature term library and clustering fraudulent and non-fraudulent annual reports. In order to establish the fraudulent feature term library, the data is preprocessed, term-pair combinations are identified, and fraudulent feature terms are filtered.

The rest of this paper is organized as follows. Section 2 reviews the process of detecting fraud for narratives in annual reports. Section 3 then develops the techniques involved in the process of detecting fraud for narratives in annual reports. Next, Section 4 demonstrates the effectiveness of the proposed fraud detection method and Section 5 provides discussion and concludes.

2. Design of a fraud detection process for narratives in annual reports

The previous section identified numerous studies that examined the use of content analysis and fraud detection of annual reports through narrative disclosures or linguistics. To enhance the detection accuracy and improve the text analysis techniques for narrative fraud in annual reports, this section proposes the process of fraud detection for narratives in annual reports, which consists of fraudulent feature term library establishment and annual report clustering, as shown in Fig. 1. Establishing fraudulent feature term library involves data preprocessing, term-pair combination, and filtering of fraudulent feature terms. Meanwhile, clustering of annual reports allows for the identification of fraudulent narratives in annual reports.

(1) Establishment of a fraudulent feature term library.

- Data preprocessing: The term set of non-fraudulent and fraudulent narratives in annual reports is extracted by using Chinese Knowledge Information Processing Group (CKIP System) (http://ckipsvr.iis.sinica.edu.tw, n.d.) for sentence breaking, part-of-speech (POS) tagging, stop-term filtering, and punctuation removal (not including comma and full stop).
- Term-pair combination: The professional terms in finance and accounting may be broken up when executing the sentence breaking for non-fraudulent and fraudulent narratives in annual reports. In this case, accurate financial and accounting terms cannot be extracted. Hence, these segmented terms must be recombined through the term-pair combination to ensure the accuracy of professional terms.
- Filtering of fraudulent feature terms: Based on the established non-fraudulent term set, fraudulent feature terms are filtered to establish a library of fraudulent feature terms in order to detect fraudulent narratives in annual reports by using the term frequency-inverse document frequency (TF-IDF) (Meijer et al., 2014; Salton and Buckley, 1988).

(2) Clustering of annual reports.

According to the established library of fraudulent feature terms, fraudulent and non-fraudulent narratives in annual reports are identified through an ensemble classifier QGA-SVM (Queen Genetic Algorithm, Support Vector Machine) that is considered the optimal prediction model for accuracy (Chen et al., 2016). These identified fraudulent and non-fraudulent narratives in annual reports are then manually confirmed with securities crime sentences, empty and misappropriation, and bounced checks of the chairman of the board (Law and Regulations Retrieving System, n.d.; Taiwan Economic Journal, n.d.) for the training dataset of fraudulent and non-fraudulent narratives in annual reports.

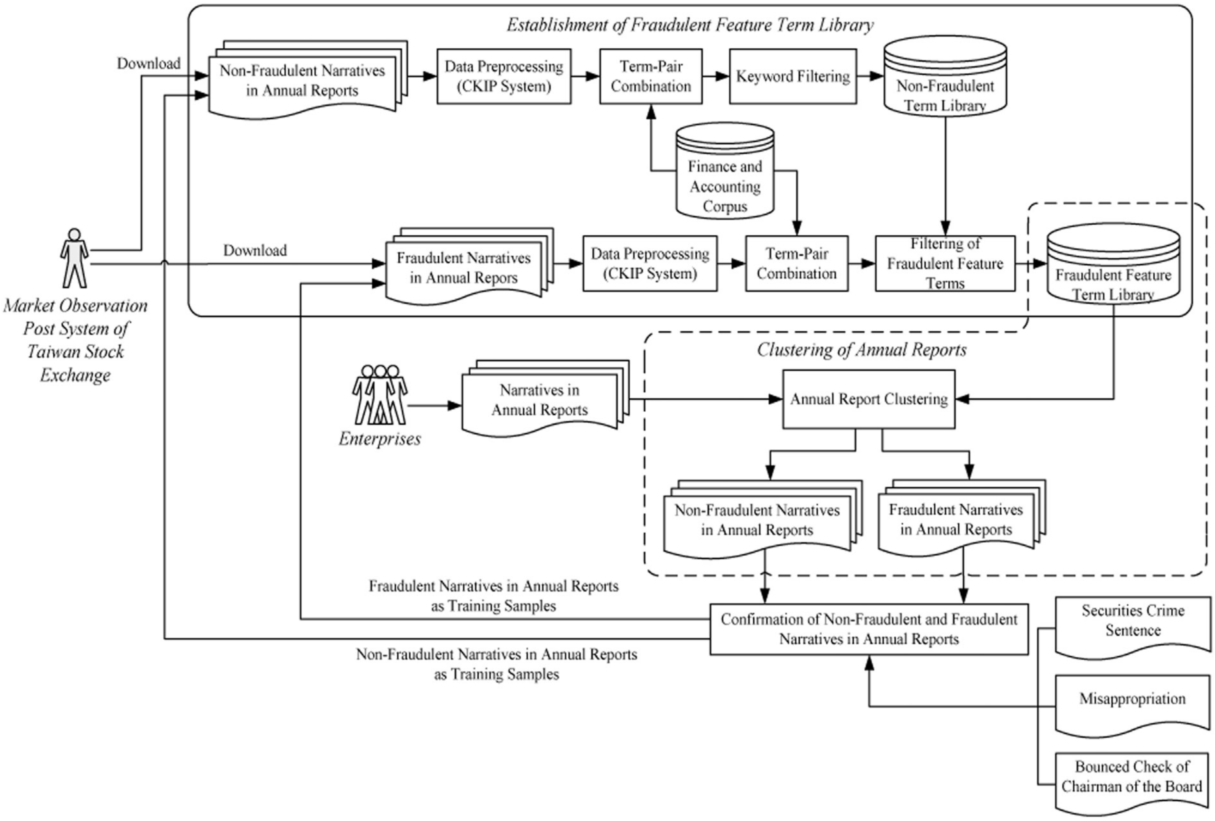
Establishment of Fraudulent Feature Term Library
Download
Non-Fraudulent Narratives in Annual Reports
Data Preprocessing (CKIP System)
Term-Pair Combination
Keyword Filtering
Non-Fraudulent Term Library
Finance and Accounting Corpus
Download
Fraudulent Narratives in Annual Repors
Data Preprocessing (CKIP System)
Term-Pair Combination
Filtering of Fraudulent Feature Terms
Fraudulent Feature Term Library
Market Observation Post System of Taiwan Stock Exchange
Clustering of Annual Reports
Enterprises
Narratives in Annual Reports
Annual Report Clustering
Non-Fraudulent Narratives in Annual Reports
Fraudulent Narratives in Annual Reports
Fraudulent Narratives in Annual Reports as Training Samples
Confirmation of Non-Fraudulent and Fraudulent Narratives in Annual Reports
Non-Fraudulent Narratives in Annual Reports as Training Samples
Securities Crime Sentence
Misappropriation
Bounced Check of Chairman of the Board

Fig. 1. Fraud detection process for narratives in annual reports.

3. Development of fraud detection techniques for narratives in annual reports

Based on the fraud detection process designed in Section 2, this section develops techniques for fraud detection, including data preprocessing, term-pair combination, filtering of fraudulent feature terms, and annual report clustering.

3.1. Data preprocessing

The CKIP (Chinese Knowledge and Information Processing) system (http://ckipsvr.iis.sinica.edu.tw, n.d.) was developed by Chinese Knowledge Information Processing Group of Institute of Information Science and the Institute of Linguistics of Academia Sinica in Taiwan. It is mainly used for Chinese natural language processing. Thus, the CKIP system is utilized in preprocessing the narratives in annual reports (e.g., report to shareholders), including segmenting sentences into meaningful terms, tagging the part-of-speech characteristics of terms, filtering stop-terms (e.g., particles and prepositions), and removing punctuations, respectively. Fig. 2 depicts the algorithm for preprocessing narratives in annual reports.

3.2. Term-pair combination

In breaking up terms from data preprocessing, professional terms in finance and accounting may be accidentally broken up, leading to incorrect professional terms. For this reason, this work designs an algorithm for term-pair combination to restore the broken up professional terms in order to facilitate the filtering of financial and accounting keywords, as depicted in Fig. 3.

3.3. Filtering of fraudulent feature terms

The term frequency-inverse document frequency (TF-IDF) (Meijer et al., 2014; Salton and Buckley, 1988) is often used as a weighting factor in information retrieval and text mining. Its value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus. To filter fraudulent feature terms in the study, the TF-IDF is thus adopted to calculate the fraudulent and non-fraudulent terms acquired from fraudulent and non-fraudulent narratives in annual reports to identify the importance of each fraudulent/non-fraudulent term for each fraudulent/non-fraudulent document. Furthermore, each fraudulent term is matched with the library of non-fraudulent term to remove non-fraudulent terms from fraudulent terms. Information gain (Quinlan, 1986; Claude, 1948) is considered as the most effective method, compared to other methods such

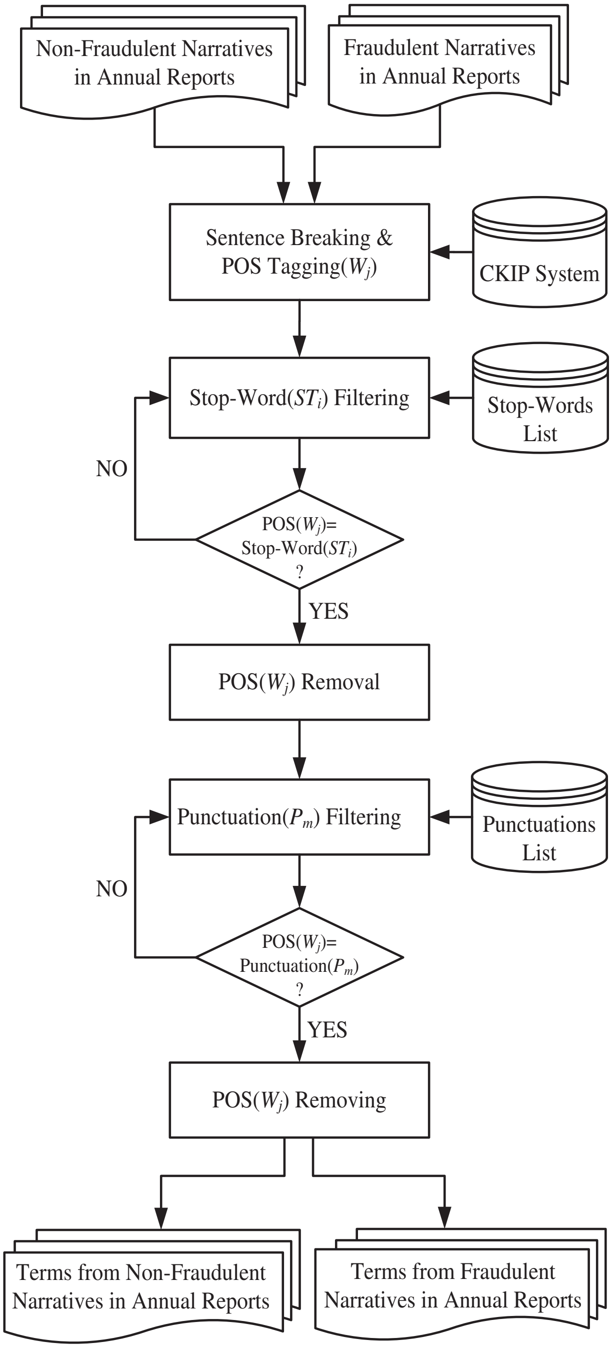
Non-Fraudulent Narratives in Annual Reports
Fraudulent Narratives in Annual Reports
Sentence Breaking & POS Tagging(W_j)
CKIP System
Stop-Word(ST_i) Filtering
Stop-Words List
NO
POS(W_j)= Stop-Word(ST_i) ?
YES
POS(W_j) Removal
Punctuation(P_m) Filtering
Punctuations List
NO
POS(W_j)= Punctuation(P_m) ?
YES
POS(W_j) Removing
Terms from Non-Fraudulent Narratives in Annual Reports
Terms from Fraudulent Narratives in Annual Reports

Fig. 2. Algorithm for preprocessing narratives in annual reports.

as term strength, mutual information, Chi Square ($\times^2$) statistic, document frequency. Based on information gain, fraudulent feature terms highly correlated with fraudulent narratives in annual reports are finally selected to establish the library for fraudulent feature terms. Fig. 4 illustrates the algorithm for filtering fraudulent feature terms, where the equations for TF-IDF and information gain are shown as Eqs. (1) and (2), respectively.

$$\mathrm{TFIDF}_{i,j} = \mathrm{TF}_{i,j} \times \mathrm{IDF}_i;\ \mathrm{TF}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}};\ \mathrm{IDF}_i = \log\left(\frac{n}{df_i}\right). \tag{1}$$

where $\mathrm{TF}_{i,j}$ is the frequency of term i appearing on a fraudulent/non-fraudulent.

document j;

IDF_t is the frequency of term i appearing on fraudulent/non-fraudulent.

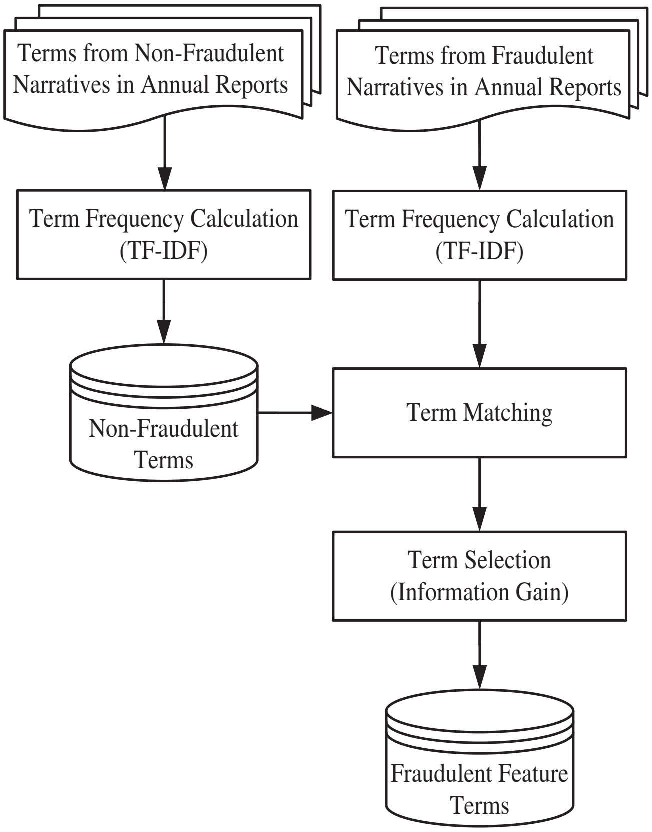
Terms from Non-Fraudulent Narratives in Annual Reports
Terms from Fraudulent Narratives in Annual Reports
Term Frequency Calculation (TF-IDF)
Term Frequency Calculation (TF-IDF)
Non-Fraudulent Terms
Term Matching
Term Selection (Information Gain)
Fraudulent Feature Terms

Fig. 3. Algorithm for term-pair combination.

documents

$n_{i,j}$ is the number of term i appearing on fraudulent/non-fraudulent.

document j;

$\sum_k n_{k,j}$ is the total number of all terms appearing on fraudulent/non-fraudulent.

documents

n is the total number of fraudulent/non-fraudulent documents;

df is the number of fraudulent/non-fraudulent documents with term i;

$$\begin{aligned} &IG(C\ |E) = H(C) - H(C\ |E) \\ &H(C) = -\sum_{i=1}^{|C|} p(c_i)\log_2 p(c_i) \\ &H(C\ |E) = \sum_{j=1}^{|E|} p(e_j) - \left[\sum_{i=1}^{|c|} p\left(c_i\ |e_j\right)\right]\log_2 p\left(c_i\ |e_j\right)\right. \end{aligned} \tag{2}$$

where $IG(C|E)$ denotes the information gain of fraudulent/non-fraudulent.

term E in fraudulent/non-fraudulent correlated term class.

C;

$H(C)$ denotes the entropy of fraudulent/non-fraudulent correlated term.

class C;

$H(C|E)$ denotes the relative entropy of fraudulent/non-fraudulent term.

Ein fraudulent/non-fraudulent correlated term class C;

$p(c_i)$denotes the probability of fraudulent/non-fraudulent correlated.

term class C;

$p(e_j)$ denotes the probability of fraudulent/non-fraudulent term E;

$p(c_i|e_j)$ denotes the probability of fraudulent/non-fraudulent term E conditional on the occurrence of fraudulent/non-fraudulent correlated term class C;

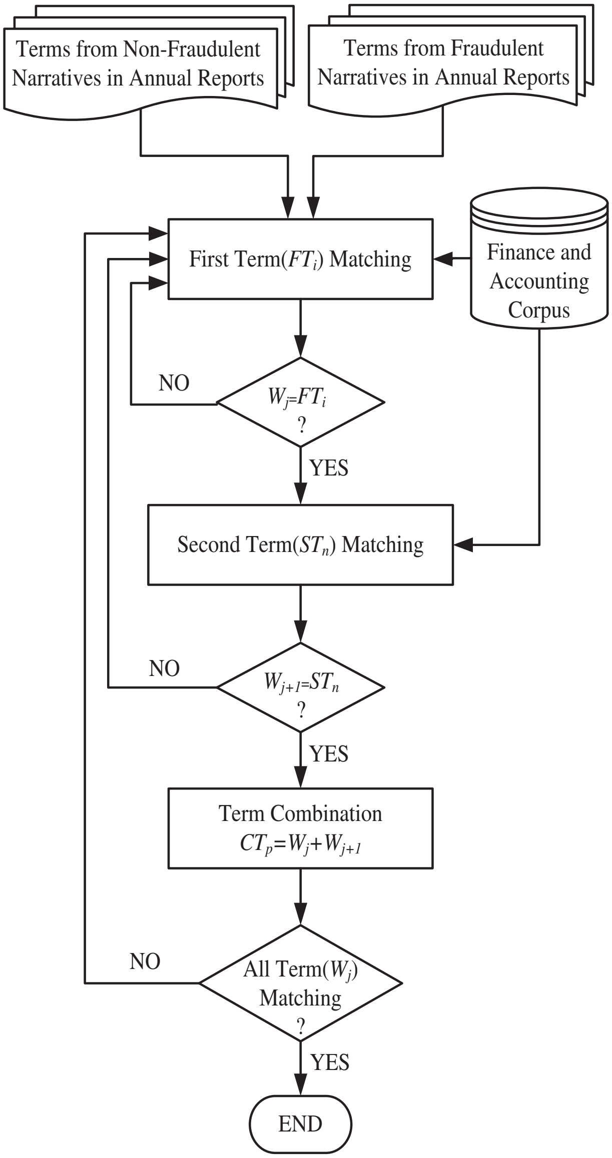
Terms from Non-Fraudulent Narratives in Annual Reports
Terms from Fraudulent Narratives in Annual Reports
First Term(FT_i) Matching
Finance and Accounting Corpus
NO
W_j=FT_i ?
YES
Second Term(ST_n) Matching
NO
W_{j+1}=ST_n ?
YES
Term Combination CT_p=W_j+W_{j+1}
NO
All Term(W_j) Matching ?
YES
END

Fig. 4. Algorithm for filtering fraudulent feature terms.

3.4. Annual report clustering

An attempt is made to accurately detect fraud in the narrative in an annual report as a valuable reference for investors, creditors, and other accounting information end users making decisions. The set of fraudulent feature terms obtained in Section 3.3 is first calculated by using the weighted method (Eq. (3)). Moreover, the weighted score is regarded as the variable value for establishing the data set. Finally, the established data set is divided into a training dataset and a testing dataset for training and testing the fraud detection model for narratives in annual reports.

$$Score_m = \frac{\sum TFIDF_{i,m}}{n_m} \tag{3}$$

where $Score_m$ represents the weighted score of the fraudulent feature term;

n_m represents the total number of words in the *m-th* article;

$TFIDF_{i,m}$ represents the product of term frequency and inverse.

document frequency of fraudulent feature term *i* appearing in.

the *m-th* article;

Support vector machine (SVM) is considered as the optimal clustering model for accuracy (Cortes and Vapnik, 1995; Oliveira and Gama, 2012; Zhou et al., 2010) and queen genetic algorithm (QGA) (Stern et al., 2006; Tsang et al., 2004) is widely used for adjusting

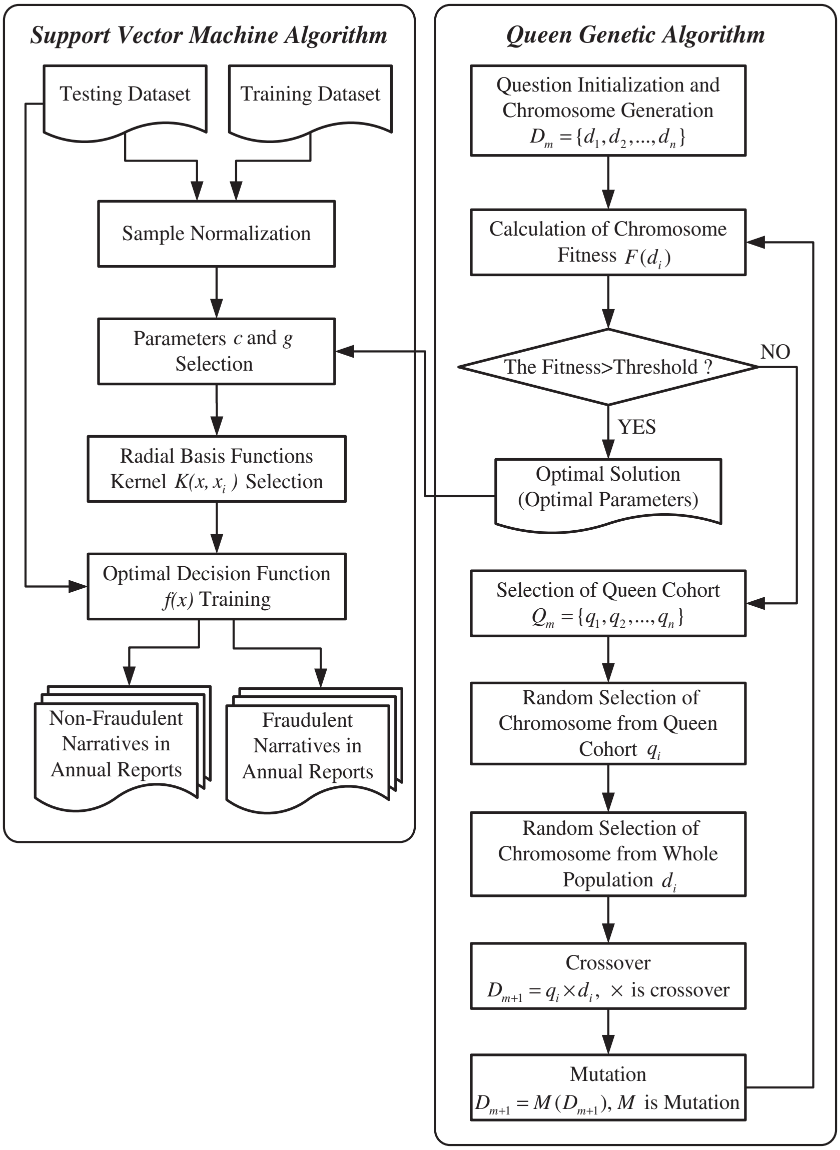
Support Vector Machine Algorithm
Testing Dataset
Training Dataset
Sample Normalization
Parameters c and g Selection
Radial Basis Functions Kernel K(x, x_i) Selection
Optimal Decision Function f(x) Training
Non-Fraudulent Narratives in Annual Reports
Fraudulent Narratives in Annual Reports
Queen Genetic Algorithm
Question Initialization and Chromosome Generation D_m = {d_1, d_2, ..., d_n}
Calculation of Chromosome Fitness F(d_i)
The Fitness>Threshold ?
NO
YES
Optimal Solution (Optimal Parameters)
Selection of Queen Cohort Q_m = {q_1, q_2, ..., q_n}
Random Selection of Chromosome from Queen Cohort q_i
Random Selection of Chromosome from Whole Population d_i
Crossover D_{m+1} = q_i × d_i, × is crossover
Mutation D_{m+1} = M(D_{m+1}), M is Mutation

Fig. 5. Algorithm for detecting fraud in annual reports.

and optimizing parameters of clustering models. Thus, this study integrates support vector machine (SVM) and queen genetic algorithm (QGA) to develop a clustering model for annual reports.

Based on the training dataset and the testing dataset established after the weighted calculation, the SVM is used for detecting fraud in annual reports and its parameters are adjusted and optimized through the QGA. Fig. 5 presents the algorithm for detecting fraud in annual reports. The related calculations are shown in Eqs. (4), (5), (6), and (7).

$$D_{m+1} = M(q_i \times d_i) \tag{4}$$

$$F(d_i) = \text{rank}(D_{m+1}) \tag{5}$$

where $F(d_i)$ denotes the fitness value;

$D_m$denotes the primal objective function;

$q_i$denotes the randomly selected fitness function in the optimal function sequence;

d_i denotes the randomly selected fitness function in all function.

Sequences;

$$f(x) = sign\left(\sum_{i=1}^{n} a_i y_i K(x, x_i) + b\right) \tag{6}$$

$$K(x, x_i) = \exp\left(\frac{\|x - x_i\|^2}{2\sigma^2}\right) \tag{7}$$

where $f(x)$ represents the optimal decision function;

a represents the Lagrange multiplier;

y represents the class index of various indicators;

b represents the offset value;

$K(x, x_i)$ represents the RBF;

σ represents the parameter of RBF;

SVM is acquired after many iteration times. The weight voting to the SVM is performed based on the weight to generate the QGA-SVM model. Eq.uation (8) shows the formula for weight voting.

$$H(x) = \arg\max \sum\nolimits_t \left(\ln \frac{1}{\beta_t}\right) h_t(x, y) \tag{8}$$

where $H(x)$ denotes the class index of QGA-SVM;

$h_t(x, y)$ denotes the class index of SVM;

β_t denotes the weight of SVM;

Finally, the testing dataset is inputted to the QGA-SVM clustering model to determine the results of annual report clustering (i.e. fraudulent narratives in annual reports or non-fraudulent narratives in annual reports).

4. Demonstration and evaluation of the proposed fraud detection method for narratives in annual reports

This section describes the fraud detection techniques for narratives in annual reports implemented using Visual Studio C#2010 and Matlab R2010b. Additionally, the feasibility and validity of the proposed method is also demonstrated using the reports to shareholders of listed companies in Taiwan. The detection accuracy is evaluated by comparing the proposed method with other fraud detection models.

4.1. Demonstration of the proposed method

This section describes the feasibility of the proposed fraud detection method for narrative annual reports, based on the reports to shareholders of listed companies in Taiwan. The detailed steps are presented as follows.

(1) Collect fraudulent/non-fraudulent narrative annual reports

In order to identify fraudulent firms, the Judicial Yuan of the Republic of China Law, as well as information published on the market observation post system, were searched for firms cited for security-related crimes. Thirty-one listed companies in Taiwan were cited for financial report fraud in 1995–2012, and these are included as the fraudulent companies. Moreover, 14 additional companies in Taiwan were classified as fraudulent on the basis of misappropriation or bounced checks as listed in the database of the Taiwan Economic Journal (TEJ). This resulted in a total of 45 fraudulent companies. The non-fraudulent companies were matched to the fraudulent companies on industry and total assets, resulting in 135 non-fraudulent companies in Taiwan.

The annual reports of the selected 45 fraudulent companies and 135 non-fraudulent companies in Taiwan were retrieved for analysis. Table 2 presents a partial report to shareholders.

(2) Preprocessing data

Step 1: Break the report into sentences and tag POS.

Through the CKIP system, the retrieved reports to shareholders (Table 2) are first broken into sentences. The part-of-speech of

Table 2
Partial report to shareholders.

The report to shareholders
本公司去年(89年度)營收實績約為70億2百萬元,稅後淨利約為7億4仟7百萬元;營收較88年度增加29億6仟1百萬元,約成長73.27%;稅後淨利較88年度增加4億6百萬元,約成長118.95%;主要係因本公司擴大掌上型電腦、ADSL及IEEE 1394系列產品的產銷規模,再加上光電事業處新增砷化鎵磊晶片生產設備順利量產,使得其效益及時顯現,因此使得營收及稅後淨利均大幅增加。(The revenue performance of the company last year (2000) was about 7 billion and 2 million dollars and the net income about 0.747 billion. The revenue was 2.961 billion dollars more than it in 1999, growing about 73.27%, and the net income increased 0.4 billion and 6 million more than it in 1999, growing about 118.95%. It was because the company expanded the production scale of palmtop computers, ADSL, and IEEE 1394 series and the production equipment for Gallium-Arsenide epitaxial wafers purchased by the office of optoelectronic business presented the mass production smoothly so that the benefits appeared in time to largely increase the revenue and net income.)

Table 3
Results of term-pair combination.

First term (*FT*)	Second term (*ST*)	Term-pair combination
稅 (taxes) (N)	後 (after) (POST)	稅後 (after taxes) (N)
稅後 (after taxes) (N)	淨利 (net income) (N)	稅後淨利 (net income after taxes) (N)
市場 (market) (N)	佔有率 (share) (N)	市場佔有率 (market share) (N)
營業 (operating) (Nv)	收入 (income) (N)	營業收入 (operating income) (N)
營運 (operational) (Nv)	目標 (goals) (N)	營運目標 (operational goals) (N)
稅後 (after taxes) (N)	虧損 (loss) (Vt)	稅後虧損 (loss after taxes) (N)
績 (accomplishment) (Vt)	效 (efficiency) (N)	績效 (performance) (N)
經營 (operation) (Vt)	績效 (performance) (N)	經營績效 (business performance) (N)
市場 (market) (N)	競爭力 (competitiveness) (N)	市場競爭力 (market competitiveness) (N)
實收 (paid) (Nv)	資本額 (capital) (N)	實收資本額 (paid-up capital) (N)

each word in these sentences is then tagged.

Step 2: Filter the stop-terms.

Based on the results of sentence breaking and part-of-speech tagging, some stop-terms are removed.

Step 3: Remove the punctuation.

Following the stop-term filtering, the punctuation is removed except for commas and full stops.

(3) Combine the term-pairs

After data preprocessing, some of the terms are combined as term-pairs. Table 3 lists those results.

(4) Filter fraudulent feature terms

Step 1: Establish the term library of fraudulent and non-fraudulent reports to shareholders.

A term library is established, based upon the terms used in the 20 fraudulent and 60 non-fraudulent reports to shareholders. This required data preprocessing and term-pair combination.

Step 2: Calculate TF-IDF.Based on the partial terms in 60 non-fraudulent and 20 fraudulent reports to shareholders, the TF-IDF of these terms is calculated based upon Eq. (1). Tables 4 and 5 list the calculation results.

Step 3: Match the terms.

The fraudulent terms listed in Table 5 are matched with the non-fraudulent terms listed in Table 4. In matching the terms, when the term appears in both Tables 4 and 5 and its TF-IDF value is larger than 2.5, the term needs to be removed from Table 5. Also, the fraudulent terms listed in Table 5 with the TF-IDF value less than or equal to 2.5 must be removed.

Step 4: Select the terms.

According to the terms acquired from step 3, the information gain is calculated by using Eq. (2) to select 242 fraudulent feature terms that have high correlations with 45 fraudulent reports to shareholders, as shown in Table 6. These 242 fraudulent feature terms are then used to cluster reports to shareholders.

Table 4
TF-IDF values for partial non-fraudulent terms.

Item no.	Term	TF value	IDF value	TF-IDF value
1	去年 (last year)	12	1.335001067	16.020011900
2	螢光粉 (fluorescent powder)	1	4.330733340	4.330733299
3	主管 (director)	2	2.538973871	5.077947617
4	廠商 (firm)	2	1.239690887	2.479381800
5	尋找 (looking for)	2	3.637586160	7.275172234
6	如 (such as)	11	0.480585739	5.286443233
7	鎖定 (lock)	2	3.637586160	7.275172234
8	加上 (plus)	10	4.330733340	43.307334900
9	第一 (first)	3	0.896746136	2.690238476
10	予以 (to be)	1	3.637586160	3.637586117
11	第三 (third)	3	2.133508763	6.400526524
12	損失 (loss)	6	2.721295428	16.32777214
13	無 (no)	1	0.804372816	0.804372787
14	精緻 (refined)	1	2.721295428	2.721295357
15	外銷 (for export)	1	3.232121052	3.232120991
16	分析 (analysis)	3	2.384823191	7.154469490
17	形象 (last year)	3	1.845826691	5.537479877
18	跨 (cross)	4	1.239690887	4.958763599
19	業外 (outside)	1	3.63758616	3.637586117
20	重大 (major)	3	2.133508763	6.400526524

Table 5
TF-IDF values for partial fraudulent terms.

Item no.	Term	TF Value	IDF Value	TF-IDF Value
1	年度 (year)	71	0.597837001	42.446426390
2	電路 (circuit)	1	2.590267165	2.590267181
3	進 (enter)	7	0.338975367	2.372827530
4	鞏固 (strengthen)	2	1.897119985	3.794239998
5	趨勢 (trend)	17	1.609437912	27.360445020
6	結合 (combine)	8	1.337504197	10.700033190
7	用高 (high)	2	4.382026635	8.764053345
8	評估 (evaluate)	3	2.995732274	8.987196922
9	看盤 (stock quote)	1	4.382026635	4.382026672
10	平衡 (balance)	5	2.772588722	13.862943650
11	179	1	4.382026635	4.382026672
12	30億 (3 billion)	1	3.688879454	3.688879490
13	逐 (gradually)	2	1.163150810	2.326301575
14	差 (difference)	1	1.437587656	1.437587619
15	鋼材 (steel)	2	4.382026635	8.764053345
16	通 (through)	2	0.531879033	1.063758016
17	基本面 (fundamental)	1	3.688879454	3.688879490
18	這 (this)	3	0.668454568	2.005363703
19	導致 (result in)	4	2.772588722	11.090354920
20	正向 (forward)	1	3.283414346	3.283414364

(5) Cluster reports to shareholders

Once all the terms have been identified, the fraud detection model needs to be trained, tested and evaluated. This is described next. The evaluation process is based upon comparing the accuracy of this fraud detection model relative to other proposed fraud detection models using the same dataset.

Step 1: Establish training and testing datasets in clustering reports to shareholders.

Following the establishment of the term library of fraudulent and non-fraudulent reports to shareholders, the residual data samples are divided into a training dataset and a testing dataset for clustering reports to shareholders (Table 7).

Step 2: Classify fraudulent/non-fraudulent reports to shareholders.

The training dataset of reports to shareholders is input into the clustering model - QGA-SVM (Fig. 5) through the use of MATLAB TOOLBOX. A ten-fold cross validation is then conducted for training and testing the clustering model. In training and testing this model, the relevant parameter settings are continuously adjusted and optimized (Chang et al., 2010; Zhou et al., 2007), as listed in Table 8. Table 9 summarizes the results of fraudulent and non-fraudulent report clustering.

4.2. Evaluation of clustering accuracy

Using the same dataset of fraudulent and non-fraudulent reports, five clustering models (i.e. Decision Tree, Grid-SVM, PSO-SVM, GA-SVM and QGA-SVM) are used to generate classification results. These classification results and the accuracy of the models are presented in Table 10. The adopted clustering model, QGA-SVM is superior to models in previous studies in terms of accuracy.

5. Conclusions

This work integrates natural language processing (NLP), queen genetic algorithm (QGA) and support vector machine (SVM) to develop a fraud detection method for narratives in annual reports. A more accurate fraud detection method should allow investors to reduce their investment risks. This research designed a process of fraud detection for narratives in annual reports. The analytical techniques related to fraud detection for narratives in annual reports are then developed and the fraud detection technique is demonstrated and evaluated.

In the experiment, the limitation of the training data set to sixty companies can be criticized. Although the statistical power is approximately 90% for the sample size, further confirmation of the discriminatory power of the proposed fraud detection method by extending the sample size would be necessary for future work. This research can also be extended to other parts of the world and other languages to see if similar results can be obtained.

The results of this research facilitate the realization of fraud detection for narratives in annual reports and the enhancement of annual report clustering accuracy to reduce investment losses and investor- and creditor-related risks, as well as enhance investment benefits.

Acknowledgements

The authors would like to thank the National Science Council of the Republic of China, Taiwan, for financially supporting this

Table 6
Fraudulent feature terms.

Item no.	Term	Item no.	Term	Item no.	Term
001	期望 (expect)	002	給予 (give)	003	合理 (reasonable)
004	機能 (function)	005	契機 (opportunity)	006	現象 (phenomenon)
007	運作 (operation)	008	誠摯 (sincere)	009	動態 (dynamic)
010	核心 (core)	011	培養 (foster)	012	建立 (establish)
013	希望 (hope)	014	利率 (interest rate)	015	好評 (praise)
016	相較 (in contrast to)	017	歷經 (go through)	018	優質 (high quality)
019	挑戰 (challenge)	020	以致 (so that)	021	開放 (open)
022	優勢 (advantage)	023	規定 (stipulate)	024	預計 (estimate)
025	面臨 (face)	026	需求 (demand)	027	致力 (dedicate)
028	多元 (multiple)	029	減少 (reduce)	030	規劃 (plan)
031	營收 (revenue)	032	密切 (close)	033	得以 (able to)
034	不佳 (poor)	035	保有 (retain)	036	顯著 (significant)
037	不如 (not as good)	038	內部 (internal)	039	評估 (evaluate)
040	預期 (anticipate)	041	進一步 (further)	042	陷入 (fall into)
043	制度 (system)	044	投資 (investment)	045	考量 (consider)
046	毛利 (gross profit)	047	朝向 (toward)	048	認列 (recognition)
049	穩健 (steady)	050	激烈 (fierce)	051	全球化 (globalization)
052	拓展 (expand)	053	股份 (shares)	054	隨時 (at any time)
055	之餘 (in addition to)	056	佔有率 (share)	057	市場 (market)
058	持平 (flat)	059	合作 cooperation)	060	效益 (benefit)
061	提高 (improve)	062	全面 (overall)	063	競爭 (competition)
064	風險 (risk)	065	基礎 (basis)	066	提昇 (boost)
067	飆漲 (soaring)	068	商機 (business opportunity)	069	衰退 (decline)
070	特性 (characteristic)	071	努力 (efforts)	072	資金 (funds)
073	比較 (compare with)	074	增進 (enhance)	075	突破 (break through)
076	利息 (interest)	077	藉由 (by)	078	採取 (adopt)
079	推動 (execute)	080	匯率 (exchange rate)	081	全球性 (global)
082	獲得 (obtain)	083	強調 (emphasize)	084	更多 (more)
085	經營 (operate)	086	以期 (hoping to)	087	審慎 (careful)
088	終於 (at last)	089	國際性 (international)	090	計畫 (program)
091	永續 (sustainable)	092	執行 (carried out)	093	受到 (suffer)
094	形成 (form)	095	全力 (all-out effort)	096	回顧 (review)
097	既有 (existing)	098	足以 (sufficient)	099	總部 (Headquarters)
100	取得 (get)	101	消費性 (consumer)	102	效果 (effect)
103	委託 (delegate)	104	提供 (provide)	105	法規 (regulations)
106	轉投資 (reinvestment)	107	成功 (success)	108	追求 (pursue)
109	盈餘 (surplus)	110	復甦 (recovery)	111	自我 (self)
112	利用 (utilize)	113	能力 (ability)	114	上下游 (upstream and downstream)
115	共同 (common)	116	預估 (forecast)	117	相當 (quite)
118	領導 (leadership)	119	精神 (spirit)	120	認證 (certification)
121	提出 (propose)	122	轉型 (transformation)	123	差異 (difference)
124	競爭力 (competitiveness)	125	夥伴 (partner)	126	事業 (cause)
127	佈局 (layout)	128	成熟 (mature)	129	謹慎 (cautious)
130	潛力 (potential)	131	價格 (price)	132	業績 (achievement)
133	顯示 (display)	134	分散 (dispersion)	135	原料 (raw material)
136	地區 (region)	137	定期 (regular)	138	定位 (position)
139	邁入 (enter)	140	證明 (prove)	141	績效 (Performance)
142	責任 (responsibility)	143	減損 (impairment)	144	多元化 (diversification)
145	訂定 (set)	146	特色 (features)	147	措施 (measures)
148	量產 (mass production)	149	完整 (complete)	150	好轉 (get better)
151	版圖 (territory)	152	時程 (schedule)	153	庫存 (stock)
154	狀況 (situation)	155	業務 (business)	156	能夠 (be able to)
157	全方位 (all-round)	158	快速 (fast)	159	面對 (face)
160	原則 (in principle)	161	未來 (future)	162	指教 (advise)
163	壓力 (pressure)	164	掌握 (grasp)	165	成立 (found)
166	獲利 (earn profits)	167	知名度 (reputation)	168	近年 (recent years)
169	致使 (to cause)	170	規模 (scale)	171	費用 (cost)
172	相對 (relatively)	173	消費 (consumption)	174	條件 (condition)
175	資本 (capital)	176	波動 (fluctuation)	177	變化 (variety)
178	維持 (maintain)	179	現金 (cash)	180	切入 (cut to)
181	尊重 (respect)	182	致 (to)	183	效能 (efficacy)
184	繼續 (carry on)	185	佈建 (construct)	186	尋求 (seek)
187	以來 (since)	188	趨於 (tend toward)	189	景氣 (boom)
190	優良 (excellent)	191	同業 (the same trade)	192	擴張 (extend)
193	應用 (application)	194	水準 (level)	195	關係 (relationship)
196	感謝 (appreciate)	197	重要 (important)	198	上漲 (rise)
199	邁進 (stride forward)	200	最佳 (optimal)	201	代表 (representative)
202	帶動 (drive)	203	困難 (difficult)	204	效率 (effectiveness)

(continued on next page)

Table 6 *(continued)*

Item no.	Term	Item no.	Term	Item no.	Term
205	估計 (estimate)	206	正式 (formal)	207	供貨 (supply goods)
208	知名 (famous)	209	理想 (ideal)	210	配合 (cooperation)
211	展望 (look into the future)	212	踏實 (pragmatic)	213	敬請 (please)
214	外部 (external)	215	重心 (focus)	216	嚴格 (strict)
217	歡迎 (welcome)	218	銷售 (sales)	219	著重 (focus)
220	平衡 (balance)	221	改善 (improve)	222	不景氣 (recession)
223	認同 (identify)	224	利益 (interests)	225	營運 (operation)
226	遭逢 (encounter)	227	秉持 (hold fast to)	228	大幅 (substantially)
229	進入 (enter)	230	高漲 (upsurge)	231	股東 (shareholder)
232	相信 (believe)	233	鼓勵 (encourage)	234	情勢 (situation)
235	低迷 (downturn)	236	衝擊 (impact)	237	達到 (achieve)
238	預測 (prediction)	239	逐漸 (gradually)	240	營業額 (turnover)
241	導致 (result in)	242	開創 (create)		

Table 7
Sample division for clustering reports to shareholders.

Sample	Fraudulent reports to shareholders	Non-fraudulent reports to shareholders
Training dataset	15	45
Testing dataset	10	30

Table 8
Parameter Settings for the QGA- SVM Model.

Parameter name	Value set
QGA population	20
QGA evolution	200
QGA threshold	0.9
c and *g* of SVM	Based on the results of QGA

Table 9
QGA-SVM testing results and detection at a significance level of 0.01.

Testing sample	Total	Correctly identified	Incorrectly identified	*P*-Value	Detected at 0.01 level	
					Upper	Lower
Fraudulent reports to shareholders	10	9	1	0.0107	10	0
Non-fraudulent reports to shareholders	30	25	5	0.0003	23	7

Table 10
Clustering accuracy comparison.

Clustering model	C	γ	Elapsed time	Accuracy
Decision tree	–	–	20.220532	75.2899%
Grid-SVM	5.3513	3.8321	19.849701	79.2632%
PSO-SVM	5.3314	5.8008	14.205772	83.8764%
GA-SVM	5.7599	6.7673	11.464359	83.2583%
QGA-SVM (used in this study)	**5.3489**	**3.8487**	**16.166594**	**85.2482%**

research under Contract Nos. NSC100-2410-H-327-003-MY2 and NSC102-2410-H-327-028-MY3. Additionally, we deeply appreciate the editor, associate editor and reviewers for their constructive comments, suggestions, and proofreading on the work.

References

Alden, Matthew E., Bryan, Daniel M., Lessley, Brenton J., Tripathy, Arindam, 2012. Detection of financial statement fraud using evolutionary algorithms. J. Emerging Technol. Account. 9 (1), 71–94.

Beattie, Vivien, McInnes, Bill, Fearnley, Stella, 2004. A methodology for analysing and evaluating narratives in annual reports: a comprehensive descriptive profile and metrics for disclosure quality attributes. Account. Forum 28 (3), 205–236.

Brazel, Joseph F., Jones, Keith L., Zimbelman, Mark F., 2009. Using nonfinancial measures to assess fraud risk. J. Account. Res. 47 (5), 1135–1166.

Breton, Gaétan, Taffler, Richard J., 2001. Accounting information and analyst stock recommendation decisions: a content analysis approach. Account. Bus. Res. 31 (2), 91–101.

Chang, Yin-Wen, Hsieh, Cho-Jui, Chang, Kai-Wei, Ringgaard, Michael, Lin, Chih-Jen, 2010. Training and testing low-degree polynomial data mappings via linear SVM. J. Mach. Learn. Res. 11, 1471–1490.

Chen, Yuh-Jen, Chen, Yuh-Min, Lu, Chang Lin, 2016. Enhancement of stock market forecasting using an improved fundamental analysis-based approach. Soft. Comput (In Press).

http://ckipsvr.iis.sinica.edu.tw/, Chinese knowledge and information processing n.d.

Churyk, Natalie Tatiana, Lee, Chih-Chen, Douglas Clinton, B., 2009. Early detection of fraud: evidence from restatements. Advances Account. Behav. Res. 12, 25–40.

Claude, E., 1948. Shannon, a mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423.

Cortes, Corinna, Vapnik, Vladimir, 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

Debreceny, Roger S., Gray, Glen L., 2010. Data mining journal entries for fraud detection: an exploratory study. Int. J. Account. Inf. Syst. 11 (3), 157–181.

Dechow, Patricia M., Ge, Weili, Larson, Chad R., Sloan, Richard G., 2011. Predicting material accounting misstatements. Contemp. Account. Res. 28 (1), 17–82.

Edward, H., 1984. Bowman, content analysis of annual reports for corporate strategy and risk. Interfaces 14 (1), 61–71.

Glancy, Fletcher H., Yadav, Surya B., 2011. A computational model for financial reporting fraud detection. Decis. Support. Syst. 50 (3), 595–601.

Gupta, Rajan, Gill, Nasib Singh, 2012. A data mining framework for prevention and detection of financial statement fraud. Int. J. Comput. Appl. 50 (8), 7–14.

Hake, Eric R., 2005. Financial illusion: accounting for profits in an Enron world. J. Econ. Issues 39 (3).

Huang, Shi-Ming, Yen, David C., Yang, Luen-Wei, Hua, Jing-Shiuan, 2008. An investigation of Zipf's law for fraud detection. Decis. Support. Syst. 46 (1), 70–83.

Humpherys, Sean L., Moffitt, Kevin C., Burns, Mary B., Burgoon, Judee K., Felix, William F., 2011. Identification of fraudulent financial statements using linguistic credibility analysis. Decis. Support. Syst. 50 (3), 585–594.

Kaminski, Kathleen A., Sterling Wetzel, T., Guan, Liming, 2004. Can financial ratios detect fraudulent financial reporting? Manag. Audit. J. 19 (1).

Kirkos, Efstathios, Spathis, Charalambos, Manolopoulos, Yannis, 2007. Data mining techniques for the detection of fraudulent financial statements. Expert Syst. Appl. 32 (4), 995–1003.

http://jirs.judicial.gov.tw/FJUD/, Law and regulations retrieving system, The Judicial Yuan of The Republic of China n.d.

Lee, Chih-Chen, Churyk, Natalie Tatiana, Douglas Clinton, B., 2013. Validating early fraud prediction using narrative disclosures. J. Forensic Investig. Account. 5 (1), 35–57.

Li, Haiying, Cai, Zhiqiang, Graesser, Arthur C., Duan, Ying, 2012. A comparative study on English and Chinese word uses with LIWC. In: Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference.

Meijer, Kevin, Frasincar, Flavius, Hogenboom, Frederik, 2014. A semantic approach for extracting domain taxonomies from text. Decis. Support. Syst. 62, 78–93.

Oliveira, Márcia, Gama, João, 2012. A framework to monitor clusters evolution applied to economy and finance problems. Intell. Data Anal. 16 (1), 93–111.

Pai, Ping-Feng, Hsu, Ming-Fu, Wang, Ming-Chieh, 2011. A support vector machine-based model for detecting top management fraud. Knowl.-Based Syst. 24 (2), 314–321.

Pennebaker, J.W., Francis, M.E., Booth, J.R., 2001. Linguistic inquiry and word count: LIWC. In: Mahwah: Lawrence Erlbaum Associates. 71.

Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., Booth, J.R., 2007. The Development and Psychometric Properties of LIWC2007. LIWC.Net, Austin, TX.

Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn. 1 (1), 81–106.

Ravisankar, P., Ravi, V., Raghava Rao, G., Bose, I., 2011. Detection of financial statement fraud and feature selection using data mining techniques. Decis. Support. Syst. 50 (2), 491–500.

Salton, Gerard, Buckley, Christopher, 1988. Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. 24, 513–523.

Siegel, Marc A., 2007. Options backdating. CPA J. 77 (19).

Skousen, Christopher J., Wright, Charlotte J., 2008. Contemporaneous risk factors and the prediction of financial statement fraud. J. Forensic Account. 9 (1), 37–61.

Stern, Helman, Chassidim, Yoash, Zofi, Moshe, 2006. Multi-agent visual area coverage using a new genetic algorithm selection scheme. Eur. J. Oper. Res. 175 (3), 1890–1907.

http://163.18.1.9/record=b1164640, Taiwan Economic Journal n.d.

Tausczik, Yla R., Pennebaker, James W., 2010. The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. 29 (1), 24–54.

Tillman, Robert, Indergaard, Michael, 2003. Pump and dump: corporate corruption in the New Economy. In: Paper presented at the American Sociological Association, Annual Meeting.

Tsang, Edward, Yung, Paul, Li, Jin, 2004. EDDIE-automation, a decision support tool for financial forecasting. Decis. Support. Syst. 37 (4), 559–565.

Wisniewski, Tomasz Piotr, Yekini, Liafisu Sina, 2015. Stock market returns and the content of annual report narratives. Account. Forum 39 (4), 281–294.

Yekini, Liafisu Sina, Wisniewski, Tomasz Piotr, Millo, Yuval, 2016. Market reaction to the positiveness of annual report narratives. Br. Account. Rev. 48 (4), 415–430.

Zhou, Lina, Burgoon, Judee K., Nunamaker, Jay F., Twitchell, Doug, 2004. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. Group Decis. Negot. 13 (1), 81–106.

Zhou, Shaowu, Wu, Lianghong, Yuan, Xiaofang, Tan, Wen, 2007. Parameters selection of SVM for function approximation based on differential evolution. In: Proceedings of the 2007 International Conference Proceedings on Intelligent Systems and Knowledge Engineering.

Zhou, Xiaofei, Jiang, Wenhan, Tian, Yingjie, Shi, Yong, 2010. Kernel subclass convex hull sample selection method for SVM on face recognition. Neurocomputing 73 (10–12), 2234–2246.