Enhancement of fraud detection for narratives in annual reports

CrossMark

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

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

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.