Financial Statement Fraud Detection using Text Mining

Rajan Gupta
Research Scholar, Department of Computer Science & Application, Maharshi Dayanand University, Rohtak, Haryana, India

Nasib Singh Gill
Professor, Department of Computer Science & Application, Maharshi Dayanand University, Rohtak, Haryana, India

***Abstract*—Data mining techniques have been used enormously by the researchers' community in detecting financial statement fraud. Most of the research in this direction has used the numbers (quantitative information) i.e. financial ratios present in the financial statements for detecting fraud. There is very little or no research on the analysis of text such as auditor's comments or notes present in published reports. In this study we propose a text mining approach for detecting financial statement fraud by analyzing the hidden clues in the qualitative information (text) present in financial statements.**

Keywords-Text Mining; Bag of words; Support Vector Machines.

I. INTRODUCTION

The illegitimate task of financial statement fraud had considerably affected the economy of a company. The analysis of financial statements assists the capital market participants in deciding about investing in a company. The information present in these statements express the performance of an organization in terms of financial status to the interested parties such as investors, creditors, auditors and management. Any deviation from Generally Accepted Accounting Principles such as presence of some extraordinary values in financial statements may results in a fraud. The presence of deviation does not always results in fraud because departures from GAAP may be appropriate to the company's situation and such departure may have been adequately disclosed.

Detection of financial statement fraud is a difficult task because of the nature of financial statements and warning signs. The mere presence of warning signs does not guarantee the occurrence of fraud and it is difficult to assess their impact before the entire fraud has unraveled. This problem is aggravated further by the fact that financial statements can be misleading even if they are in accordance with GAAP.

Financial statements released by companies consist of textual information in form of auditor's comments and disclosure as footnotes along with financial ratios. This qualitative information may contain indicators of fraudulent financial reporting in form of strategically placed phrases. In order to conceal the fraudulent activity, perpetrators may use selective sentence constructions, selective adjectives and adverbial phrases. Financial statement fraud can be detected by analyzing the above mentioned signals hidden in textual information present in published financial reports.

Companies may present a rosy picture to the investors by manipulating the financial measurements and qualitative narratives of financial statements. These disclosures (qualitative narratives) may not contain fraud indicators explicitly; however indicators of fraud can be constructed by understanding the syntactic as well as semantics of any natural language because perpetrators of fraud may camouflage the indicators by using semantic arsenal of the language. Therefore, in order to detect fraud, it is necessary to examine the qualitative disclosures in the footnotes in the financial statements, as well as the numbers (quantitative information) associated with financial statements.

Quantitative information has been analyzed by number of researchers for detection of fraudulent financial reporting. Therefore, in order to detect fraud indicators present in qualitative contents of financial statements, we present a text mining approach for differentiating between fraud and non – fraud financial statements.

The textual information present in financial statements is unstructured in nature. Text is generally amorphous and therefore must be converted into structured data before applying any predictive data mining techniques such as classification or unsupervised learning method such as clustering in order to detect fraudulent financial reporting.

Text mining is a process of extracting meaningful numeric indices (structured data) from unstructured text. Text mining can analyze words or cluster of words and can be used for determining the relationship with other variables of interest such as fraud or non fraud. Therefore, a text mining approach for detecting fraudulent financial reporting is presented in this paper. The rest of the paper is organized as follows. Section 2 presents a brief overview of the research done in the field of detection of financial statement fraud and identifies the need of an approach for analyzing text present in financial statements for detecting fraudulent financial reporting. Section 3 represents a text mining approach for detection of financial statement fraud followed by conclusion (Section 4).

II. LITERATURE REVIEW

A number of researchers have devoted a significant amount of effort in detecting fraudulent financial reporting. In order to detect fraud several researchers have used various data mining techniques.

For instance, Koh and Low [1] constructed a decision tree by using a data sample of 165 organizations. In order to detect fraud, following six financial variables were examined: quick assets to current liabilities, market value of equity to total assets, total liabilities to total assets, interest payments to earnings before interest and tax, net income to total assets, and retained earnings to total assets. Cecchini M. [2] in 2005 examined quantitative variables along with text information for detection of fraud. The qualitative variables were mapped to a higher dimension which takes in to account ratios and year over year changes.

Kotsiantis et al [3] explored the effectiveness of machine learning techniques such as Decision Tree, Artificial Neural Network, Bayesian Network, K – Nearest Neighbour, Support Vector Machines in detecting firms that issue fraudulent financial statements. The 41 fraudulent firms were matched with 123 non- fraudulent firms. All the variables used in the sample were extracted from formal financial statements, such as balance sheets and income statements.

In 2007, Kirkos et al [4] investigated the usefulness of three Data Mining classification methods namely Decision Trees, Neural Networks and Bayesian Belief Networks by analyzing 27 financial ratios extracted from publicly available data of 76 Greek manufacturing firms for detecting fraudulent financial statements. Further, Hoogs et al [5] developed a genetic algorithm approach for detecting financial statement fraud by analyzing 76 comparative metrics, based on specific financial metrics and ratios that capture company performance.

Belinna et al [6] examined the effectiveness of CART on identification and detection of financial statement fraud by analyzing financial ratios from financial reports of 148 organizations and found CART as a very effective technique in classifying financial statements as fraudulent or non – fraudulent.

Ibrahim et al [7] examined the efficiency of data mining techniques i.e. decision tree and neural network for detection of financial statement fraud by analyzing data from 100 manufacturing firms and concluded that leverage ratio and return on assets ratios are important financial ratios in detecting financial statement fraud.

Furthermore, Ravishankar et al [8] in 2011 applied six data mining techniques namely Multilayer Feed Forward Neural Network (MLFF), Support Vector Machines (SVM), Genetic Programming (GP), Group Method of Data Handling (GMDH), Logistic Regression (LR), and Probabilistic Neural Network (PNN) to identify companies that resort to financial statement fraud on a data set obtained from 202 Chinese companies of which 101 were fraudulent and 101 were non-fraudulent companies. The input vector used by them consists of 35 financial variables or ratios extracted from publically available financial statements.

Recently, Gupta et al [9] examined the efficacy of three data mining techniques namely CART, Naïve Bayesian Classifier and Genetic Programming for detecting financial statement fraud by analyzing 52 financial ratios extracted from financial statements of 114 organizations.

The review of existing academic literature reveals that research conducted till date in the field of detection of financial statement fraud had majorly analyzed financial ratios or variables which can be extracted from financial statements. A very few studies have analyzed the key component of financial statements i.e. qualitative contents in order to detect fraud.

In order to detect hidden valuable knowledge from textual financial data, we propose a text mining approach in this study because traditional mining techniques are insufficient in detecting fraud from the increasing amount of text data.

III. TEXT MINING: AN APPROACH FOR DETECTION OF FINANCIAL STATEMENT FRAUD

Figure 1 illustrates the proposed text mining approach for financial statement fraud detection. Text mining system takes as an input the collection of financial statements. In order to detect fraudulent financial reporting, financial statements of both type of organizations (fraudulent or non fraudulent) need to be collected as the first step. Companies with fraudulent history can be identified by analyzing AAER's issued by SEC. Data set should contain financial statements of non fraud organization for each fraudulent organization. The non fraud organization should be of same size (on the basis of assets or sales) as that of fraudulent organizations.

Second step is preprocessing which involves the extraction of qualitative narratives from financial statements and arranging into a document because a document is a basic unit of analysis in text mining. During preprocessing, words present in all the documents should be converted into lower case so as to avoid inclusion of two same words such as "Legal" and "legal" as different words in the corpus (collection of documents).

All the punctuations should be removed from the corpus followed by removal of any number if present because input to the classifiers should contain only text. Stopwords such as articles (a, the etc.), conjunctions (but, and etc.) and prepositions (on, in etc.) should also be removed during preprocessing because these words does not help in discriminating the documents. Stemming is not required in domain of accounts because inflected terms may have different meanings.

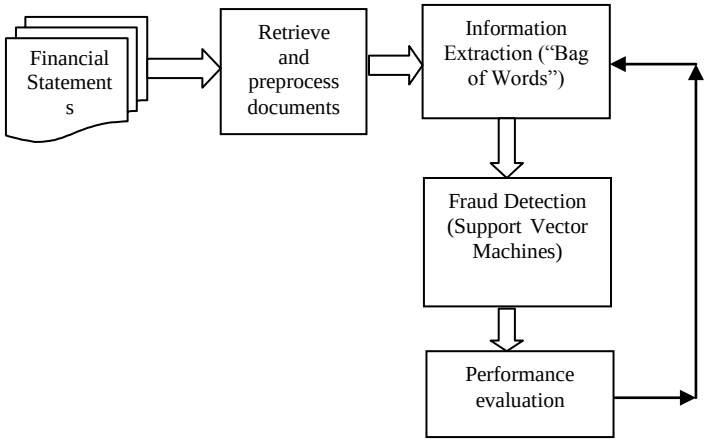

Figure 1: Text Mining detection for financial statement fraud

Since, in text mining, a sentence is regarded as a set of words and order of words can be changed with no impact on the result of the analysis, therefore syntactical structure of a sentence can be ignored for handling the text in an efficient manner. However, information regarding number of occurrences of each word should be retained. This unordered collection of words is known as "bag of words". In "bag of words" approach, the occurrence of each word is used as a feature for training a classifier. The "bag of words" model represents each document with a vector of word count that appears in the document. The vector associated with each document is compared with typical vector associated with a given class (fraud or non fraud). Documents with similar vectors are considered to be similar in content and dissimilar otherwise.

The vector spaces generated above will be used by next step for classifying organizations into fraud or non fraud. We recommend the use of Support Vector Machine – a supervised classification method, for detecting fraudulent financial reporting because SVM's construct a hyperplane in feature space which best classifies among fraudulent or non fraudulent financial reporting. SVM takes a set of input data and predicts, for each given input, which of two possible classes (fraud or non fraud) forms the output. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other.

Since SVM is a supervised machine learning method, it will learn from feature spaces of both fraudulent and non fraudulent examples present in the training set. After learning, this method is capable of classifying correctly between fraud and non fraud organizations present in the testing dataset. The accurateness of classification should be evaluated by using evaluation measures such as accuracy, precision, recall (sensitivity in binary classification), F-measure and purity.

IV. CONCLUSION

In this conceptual paper, we presented a text mining approach for detection of financial statement fraud. Fraud detection model presented in this paper begins with collection of financial statements for both fraud and non fraud organizations followed by preprocessing which involves lexical analysis of text present in financial statements. At the next step, bag of words approach has been selected for extracting information hidden in the text which results in vector spaces for both fraudulent and non fraudulent organizations.

These vector spaces acts as an input vector to the Support vector machines which learns from training data and further classifies organizations from testing data into fraud or non fraud. Finally, the correctness of classification is measured by using standard evaluation measures.

The methodology proposed in this paper for detection of financial statement fraud differs from earlier methodology in terms of input vector. Input vector in most of the previous studies consists of financial ratios and metrics i.e. quantitative information present in financial statements. Unlike earlier research studies, we selected text i.e. qualitative narratives present in financial statements in order to assess likelihood of financial statement fraud.

Financial statement fraud is a major concern for most of the organization worldwide. Hence both the quantitative and qualitative information available in annual reports should be analyzed simultaneously for assessing the risk of fraud.

REFERENCES

[1] H.C. Koh, C.K. Low, Going concern prediction using data mining techniques, Managerial Auditing Journal 19 (3) (2004) 462–476.

[2] Cecchini M. 2005. Quantifying the risk of financial events using kernel methods and information retrieval. Doctoral dissertation, University of Florida.

[3] Kotsiantis S., Koumanakos E., Tzelepis D. and Tampakas V. "Forecasting Fraudulent Financial Statements using Data Mining", International Journal of Computational Intelligence VOLUME 3 NUMBER 2 2006.

[4] Efstathios Kirkos, Charalambos Spathis &Yannis Manolopoulos (2007), Data mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications 32 (23) (2007) 995–1003.

[5] Hoogs Bethany, Thomas Kiehl, Christina Lacomb and DenizSenturk (2007). A Genetic Algorithm Approach to Detecting Temporal Patterns Indicative Of Financial Statement Fraud, Intelligent systems in accounting finance and management 2007; 15: 41 – 56, John Wiley & Sons, USA, available at: www.interscience.wiley.com.

[6] BelinnaBai, Jerome yen, Xiaoguang Yang, False Financial Statements: Characteristics of china listed companies and CART Detection Approach, International Journal of Information Technology and Decision Making , Vol. 7, No. 2(2008), 339 – 359.

[7] Ibrahim H. , Ali H. "The use of data mining techniques in detecting fraudulent financial statements: An application on manufacturing firms", The journal of faculty of economics and administrative sciences, (2009) Vol. 14, No. 2 pp. 157 – 170.

[8] P.Ravisankar, V. Ravi, G.RaghavaRao, I., Bose, Detection of financial statement fraud and feature selection using data mining techniques, Decision Support Systems, 50(2011) 491 – 500.

[9] Gupta Rajan, Gill N.S. 2012 "Data Mining Techniques – A Key for detection of financial statement fraud" , International Journal of Computer Science and Information Security, Volume 10 No. 3, pp. 49 – 57.