Due to the nature of financial statement fraud, one of the most difficult tasks in detecting such  
fraud is the identification of its symptoms. Some of those symptoms might be present even though  
no fraud exists. Generally Accepted Accounting Principles GAAP violations, for example, may  
not necessarily indicate presence of fraud since departures from GAAP may be appropriate to the  
company’s situation and such departures may have been adequately disclosed. It is possible that  
only a small number of symptoms may manifest themselves when fraud is occurring because, for  
example, the fraud may be related to omission by management of disclosures on contingent  
liabilities or related-party transactions from the notes to the financial statements, and it is difficult  
to assess their impact before the entire fraud has unraveled. Since symptoms of fraud can be  
caused by legitimate factors, the mere presence of symptoms cannot necessarily lead to inference  
of fraud. Moreover, fraud symptoms cannot easily be ranked in order of importance, nor can they  
be easily combined to yield effective predictive models. Their relative importance varies widely.  
Fraud detection is muddled by the lack of consensus on symptoms that reliably indicate fraudulent  
behavior. Nevertheless, it is widely acknowledged that fraud symptoms often exhibit themselves  
through changes in the financial statements.  
The difficulty of detecting fraud is further exacerbated by the fact that financial statements can  
be misleading even if they are in conformity with GAAP. This is due to the fact that the U.S.  
GAAP is rules-based and rules cannot be complete in the sense of covering all conceivable  
situations. It is possible for the companies to be creative in financial measurements as well as  
disclosures. Therefore, it is necessary to investigate the quantitative information in the financial  
statements, as well as the qualitative disclosures in the footnotes accompanying the financial  
statements.

Until recently, most researchers have modeled fraud detection by traditional statistical techniques such as logistic regression Persons 1995; Beasley 1996; Summers and Sweeney 1998; Lee  
et al. 1999; Abbot et al. 2000; Bell and Carcello 2000; Spathis 2002, linear discriminant analysis  
Fanning and Cogger 1998; Kaminski et al. 2004, and probit analysis Dopuch et al. 1987;  
Hansen et al. 1996; Beneish 1999; Lennox 2000. More recently, the studies have used data  
mining and machine learning techniques to model problems in the domains of accounting and  
finance. This shift can be attributed to the limitations of the traditional statistical techniques used  
in the earlier studies. Drawing on the field of Artificial Intelligence AI, some of the fraud  
detection models have used neural networks Green and Choi 1997; Fanning and Cogger 1998,  
expert systems Ragothaman et al. 1995; Eining et al. 1997, genetic algorithms Hoogs et al.  
2007, and decision trees Kirkos et al. 2007 to detect fraud.  
A perusal of the above literature shows that most of the studies used financial metrics and  
ratios extracted from financial statements to detect fraud. Some of these studies have focused on  
examining the relationship between fraudulent financial reporting and quantitative indicators such  
as composition of boards of directors, insider trading, auditor rotation, or financial restatements, in  
addition to financial data.  
Furthermore, it should be noted that many studies, including Hansen et al. 1996, Eining et  
al. 1997, and Bell and Carcello 2000, used internally generated financial information. On the  
other hand, fraud studies by researchers such as Green and Choi 1997, Summers and Sweeney

1998, Beneish 1999, Kaminski et al. 2004, Hoogs et al. 2007, and Kirkos et al. 2007  
showed the benefits of using external information. Summers and Sweeney 1998 demonstrated  
that their findings hold even when fraud risk factors from prior studies were controlled, indicating  
an incremental benefit to using external information.  
However, the limitations of these models to correctly predict fraud can have serious implications due to high rates of false negatives Type I error and false positives Type II error.  
Typically, the cost of misclassifying a company involving fraud i.e., a false negative is higher  
than the cost of misclassifying a no-fraud company i.e., a false positive. For example, if an  
investor invests in a company that is involved in fraud, but this company has been misclassified as  
a no-fraud company, he will incur a loss when fraud is discovered. On the other hand, if he does  
not invest in a no-fraud company as this company is misclassified as a fraud company, he will miss  
a profitable investment opportunity.  
Kaminski et al. 2004 demonstrated the limited ability of financial ratios to detect fraud and  
concluded that these conventional quantitative financial factors are inadequate for predicting fraud.  
More recently, Dikmen and Küçükkocaolu 2010 have used a sample of 126 Turkish manufacturing firms described over ten financial ratios to detect factors associated with false financial  
statements with 82 percent accuracy. Dechow et al. 2011 conducted a detailed analysis of firms  
investigated by the SEC for misstating quarterly or annual earnings. Using F-ratio, they predicted  
fraud with 79 percent accuracy.  
In contrast, in this paper we use the verbal, qualitative nonquantitative content of the annual  
reports to build our fraud detection model, as textual content of annual reports contains richer  
information than the financial ratios, which can be easily camouflaged. As our results show, our  
model performs better than the earlier fraud detection models see the “Results and Discussion”  
section

