Qualitative information in annual reports & the detection of corporate fraud: a natural language processing perspective

Sunita Goel
University at Albany, State University of New York, sgoels@hotmail.com

The University at Albany community has made this article openly available. Please share how this access benefits you.

Follow this and additional works at: https://scholarsarchive.library.albany.edu/legacy-etd

Part of the Accounting Commons, Library and Information Science Commons, and the Linguistics Commons

Recommended Citation

This Dissertation is brought to you for free and open access by the The Graduate School at Scholars Archive. It has been accepted for inclusion in Legacy Theses & Dissertations (2009 - 2024) by an authorized administrator of Scholars Archive. Please see Terms of Use. For more information, please contact scholarsarchive@albany.edu.
Qualitative Information in Annual Reports

& The Detection of Corporate Fraud:

A Natural Language Processing Perspective

by

Sunita Goel

A Dissertation

Submitted to the University at Albany, State University of New York

in Partial Fulfillment of

the Requirements for the Degree of

Doctor of Philosophy

College of Computing and Information

Department of Informatics

2009
Qualitative Information in Annual Reports & The Detection of Corporate Fraud: A Natural Language Processing Perspective

by

Sunita Goel

COPYRIGHT 2009
ABSTRACT

High profile cases of fraudulent financial reporting such as those that occurred at Enron and WorldCom have shaken public confidence in the U.S. financial reporting process and have raised serious concerns about the roles of auditors, regulators, and analysts in financial reporting. In order to address these concerns and restore public confidence, the Sarbanes-Oxley Act (SOX) of 2002 was enacted. However, SOX has not lived up to its promise. Numerous cases of fraudulent financial reporting have surfaced in the post-SOX era. So far, the major thrust of research has been on examining fraud that has already been discovered. This dissertation creates a methodology to proactively identify means to detect fraud by examining the qualitative content of annual reports using natural language processing tools. The methodology is created using Support Vector Machines, a supervised machine learning technique. In this research, we examine both the verbal content and the presentation style of the qualitative portion of the annual reports and seek to explore linguistic features that distinguish fraudulent annual reports from non-fraudulent annual reports. To detect fraud, it is important to investigate qualitative content as textual content of annual reports contains richer information than the financial ratios, which can be easily camouflaged. This study also creates a classification metric for early prediction of fraud by examining changes in the qualitative content of annual reports for pre-fraud, fraud and post-fraud periods of fraud companies. What distinguishes this methodology from earlier research on fraud detection is its use of
qualitative textual content in annual reports as opposed to quantitative financial information such as ratios, which have limited ability to predict fraud as discussed in the literature. Our results indicate that employment of linguistic features is an effective means to detect fraud.
ACKNOWLEDGMENTS

I am forever grateful to my husband for his unconditional love and support throughout my dissertation. I am also thankful to my son, Samir, for his patience and understanding. He seldom complained when I was not around and was spending long hours at the library carrel, immersed in my research. Without the moral support of my husband, love of my son Samir, and blessings of our parents, this dissertation would not have been possible. I also would like to thank my sister, Minu, and brother, Himanshu, for keeping me entertained with their amusing stories during stressful times. This thesis is dedicated to my family who were with me all the way in this difficult journey.

I am fortunate for having an exceptional doctoral committee. I would like to thank Dr. Jagdish Gangolly for getting me started with this research and for his enthusiasm for linguistics. Thanks for always believing in me and encouraging me to think independently. I also wish to thank Dr. Sue Faerman for teaching me the fundamentals of research and scientific writing and Dr. Ozlem Uzuner for challenging me and not letting me settle for anything less than the best. I am grateful to my committee for sharing their knowledge and expertise with me.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>v</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>ix</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xii</td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 MOTIVATION</td>
<td>4</td>
</tr>
<tr>
<td>1.2 PROBLEM STATEMENT</td>
<td>7</td>
</tr>
<tr>
<td>1.3 IMPORTANCE OF THIS RESEARCH</td>
<td>9</td>
</tr>
<tr>
<td>1.4 BRIEF OVERVIEW OF RESEARCH METHODOLOGY</td>
<td>12</td>
</tr>
<tr>
<td>1.5 OVERVIEW OF THE THESIS</td>
<td>14</td>
</tr>
<tr>
<td>2 LITERATURE REVIEW</td>
<td>17</td>
</tr>
<tr>
<td>2.1 APPROACHES TO FRAUD DETECTION USING AI AND NON-AI TECHNIQUES</td>
<td>17</td>
</tr>
<tr>
<td>2.2 QUALITATIVE ANALYSIS OF COMPANY REPORTS</td>
<td>32</td>
</tr>
<tr>
<td>2.3 COMPARISON WITH RELATED WORK</td>
<td>39</td>
</tr>
<tr>
<td>2.4 SUMMARY</td>
<td>41</td>
</tr>
<tr>
<td>3 METHODOLOGY</td>
<td>43</td>
</tr>
<tr>
<td>3.1 FRAUD CLASSIFICATION PROBLEM</td>
<td>43</td>
</tr>
<tr>
<td>3.2 CHOICE OF CLASSIFIER – SVM</td>
<td>46</td>
</tr>
<tr>
<td>3.3 FEATURE SELECTION</td>
<td>48</td>
</tr>
<tr>
<td>3.4 HYPOTHESES AND VARIABLES</td>
<td>51</td>
</tr>
<tr>
<td>3.4.1 HYPOTHESES</td>
<td>52</td>
</tr>
<tr>
<td>3.4.2 PREDICTOR VARIABLES</td>
<td>55</td>
</tr>
<tr>
<td>3.4.3 DEPENDENT VARIABLE</td>
<td>59</td>
</tr>
<tr>
<td>3.4.4 CONTROL VARIABLE</td>
<td>59</td>
</tr>
<tr>
<td>3.5 PROCEDURE OF HYPOTHESIS TESTING</td>
<td>60</td>
</tr>
<tr>
<td>3.6 SAMPLE AND DATA COLLECTION</td>
<td>63</td>
</tr>
<tr>
<td>3.7 PREPROCESSING OF THE DATASET</td>
<td>66</td>
</tr>
<tr>
<td>3.8 TRAINING THE CLASSIFIER</td>
<td>67</td>
</tr>
<tr>
<td>3.9 TESTING THE CLASSIFIER</td>
<td>69</td>
</tr>
</tbody>
</table>
LIST OF TABLES

TABLE 1.1 ELEMENTS OF FINANCIAL STATEMENT FRAUD.................................................................3

TABLE 3.1 TEMPLATE OF A 2X2 CONTINGENCY TABLE FOR TESTING A FRAUD HYPOTHESIS..........................................................................................................................61

TABLE 3.2 TEMPLATE OF A 2X2 CONTINGENCY TABLE FOR TESTING A FRAUD HYPOTHESIS RELATING TO VOICE..............................................................................................................61

TABLE 4.1 SAMPLE SELECTION........................................................................................................78

TABLE 4.2 DISTRIBUTION OF FRAUD AND NO-FRAUD COMPANIES BY INDUSTRIAL SECTOR.................................................................................................................................80

TABLE 4.3 SAMPLE FOR FRAUD DETECTION - VERSION 1...............................................................82

TABLE 4.4 SAMPLE FOR FRAUD DETECTION - VERSION 2...............................................................82

TABLE 4.5 SAMPLE FOR DETECTING EARLY WARNING SIGNS OF FRAUD........................................82

TABLE 5.1 DIFFERENT KERNELS AVAILABLE TO TRAIN SVMS.....................................................102

TABLE 5.2 TRAINING AND TESTING DATASETS.............................................................................104

TABLE 5.3 BASELINE RESULTS WITH NB CLASSIFIER.................................................................105

TABLE 5.4 RANKING OF THE TOP 78 DISCRIMINATIVE WORDS BY INFORMATION GAIN FOR DETECTING FRAUD..................................................................................................108

TABLE 5.5 RANKING OF THE TOP 73 DISCRIMINATIVE WORDS BY INFORMATION GAIN FOR PREDICTING LEVELS OF FRAUD...............................................................................109

TABLE 5.6 BASELINE RESULTS WITH SVM CLASSIFIER.............................................................113

TABLE 5.7 TEMPLATE OF A CONFUSION MATRIX TABLE..............................................................114
TABLE 7.2 CONTINGENCY TABLE FOR TESTING A FRAUD HYPOTHESIS RELATING TO TONE: VERSION 2...........................................................................................................163
TABLE 7.3 CHI-SQUARE RESULTS FOR THE TONE CATEGORY WORDS.................................163
TABLE 7.4 CONTINGENCY TABLE FOR TESTING A FRAUD HYPOTHESIS RELATING TO VOICE: VERSION 1...........................................................................................................165
TABLE 7.5 CONTINGENCY TABLE FOR TESTING A FRAUD HYPOTHESIS RELATING TO VOICE: VERSION 2...........................................................................................................166
TABLE 7.6 CHI-SQUARE RESULTS FOR USE OF PASSIVE VOICE AND ACTIVE VOICE SENTENCES........................................................................................................................166
TABLE 7.7 CONTINGENCY TABLE FOR TESTING A FRAUD HYPOTHESIS RELATING TO UNCERTAINTY MARKERS: VERSION 1.................................................................167
TABLE 7.8 CONTINGENCY TABLE FOR TESTING A FRAUD HYPOTHESIS RELATING TO UNCERTAINTY MARKERS: VERSION 2.................................................................168
TABLE 7.9 CHI-SQUARE RESULTS FOR USE OF UNCERTAINTY MARKERS..........................168
LIST OF FIGURES

FIGURE 1.1 FINANCIAL STATEMENT FRAUD AAERS ISSUED BY THE SEC, BY YEAR OF ISSUE.........................................................................................................................................................10

FIGURE 4.1 DISTRIBUTION OF FRAUD COMPANIES BY LAST YEAR OF FRAUD PERIOD........................................................................................................................................................................79

FIGURE 5.1 ILLUSTRATION OF “BAG OF WORDS” APPROACH........................................88

FIGURE 5.2 CASE OF PERFECT CLASSIFICATION.................................................................................................................................93

FIGURE 5.3(a) HYPERPLANE WITH SMALLEST MARGIN..................................................96

FIGURE 5.3(b) HYPERPLANE WITH WIDEST MARGIN...................................................96

FIGURE 5.4 CASE OF NONLINEAR CLASSIFICATION....................................................97

FIGURE 5.5 SAMPLE OF ARFF FILE USED IN WEKA..................................................112

FIGURE 5.6 SNAPSHOT OF ATTRIBUTES VISUALIZED IN WEKA .................................119
CHAPTER 1
INTRODUCTION

Corporate reports are important vehicles that companies use to communicate information on past year performance as well as prospects of future performance to all interested parties. In the United States (U.S.), to comply with the securities laws, companies whose stock is traded on the US stock exchanges, unless they are exempt from the reporting requirements, must file with the Securities and Exchange Commission (SEC) such annual reports prepared in conformity with the Generally Accepted Accounting Principles (GAAP) and audited by accountants registered with the SEC. GAAP is a framework of guidelines and includes standards, conventions, and rules that must be followed in the preparation of financial disclosures in the corporate financial reports. The financial statements and the disclosures contained in such corporate financial reports must be free of material errors, must not be misleading, and must include disclosures relating to all material events affecting the financial condition of the company.

For the purpose of this dissertation, the term “financial statement fraud” refers to fraud committed when the financial reports contain misrepresentation of material facts or such material facts have not been adequately disclosed in such reports. Since fraud involves deliberate wrongful acts performed with an intent to gain unfair advantage by deceiving another person, financial statement fraud involves intentional misstatements or omission of material facts affecting the financial condition of the companies.
Fraud typically exhibits three characteristics: motive, opportunity, and rationalization, usually referred to as the “fraud triangle” (Cressey, 1950). These three characteristics can be thought of as preconditions for fraud. Montogomery et al. (2002) explain these three characteristics of the fraud triangle in the context of financial statement fraud as follows: for fraud to occur, first, there must be incentives or pressures to materially misstate or to omit material facts from the financial statements; second, there must be opportunities to carry out such material misstatements or omissions; and third, there must be values, beliefs and attitudes that allow one to knowingly and intentionally commit a dishonest act, and rationalize committing such a dishonest act.

Elliot and Willingham (1980) define financial statement fraud as deliberate fraud committed by top management that injures users of financial statements such as investors and creditors through materially misleading financial statements. The common themes among different definitions of financial statement fraud that have been provided in the literature (Sawyer, 1988; Thornhill and Wells, 1993; Arens and Loebbecke, 1994; Vanasco, 1998; Albrecht et al., 2001) and in official pronouncements by authoritative bodies (Institute of Internal Auditors, 1985, 1986; National Commission on Fraudulent Financial Reporting, 1987; Association of Certified Fraud Examiners, 1993, 1996) are summarized in Table 1.1.
TABLE 1.1

*Elements of Financial Statement Fraud*

- Intentional conduct, whether by act or omission
- Committed by management
- Results in materially misleading financial statements (which may arise from misrepresentation or omission of material facts)
- Concealment through fraudulent financial reporting (perpetrators have taken steps to hide fraud from others)
- Users of financial statements have relied and acted upon them and in the process have been injured

Financial statement fraud is defined in different ways, but it can be summarized as an illegitimate act, committed by management, which injures other parties through misleading financial statements. In the literature, the terms “financial statement fraud” and “management fraud” have been used interchangeably. This is primarily due to the fact that management is responsible for producing reliable financial statements that are free of any material misstatement, error, irregularity or fraud and, when financial statement fraud occurs, it typically is with consent or knowledge of management (Elliott and Willingham, 1980; Robertson, 2000).

In this study, fraud\(^1\) is construed to be consistent with how financial statement fraud or management fraud is defined in the literature. The remainder of this chapter is organized as follows. The first two sections of this chapter provide motivation for the issues

---

\(^1\) The terms fraud, corporate fraud, and financial statement fraud are used interchangeably for rest of this dissertation.
considered in this dissertation and state the problem considered. The third section explains why the research problem addressed in this dissertation is important. This chapter also presents a brief overview of the proposed research methodology, and concludes with an outline of the structure of this dissertation.

1.1 MOTIVATION

A series of recent corporate frauds resulting from fraudulent financial reporting, particularly on the scale that occurred at Enron, WorldCom, and Tyco, have evoked visceral reactions from investors, financial analysts, auditors, regulators, employees, and the American public in general. As a result of these accounting scandals, investor confidence in the U.S. financial reporting process and the reliability of financial statements has greatly eroded.

The Auditing Standards Board Statement on Auditing Standards (SAS) NO. 99, *Consideration of Fraud in a Financial Statement Audit*, requires auditors to provide reasonable assurance that the audited financial statements are free of material fraud. In cases where fraud is detected, the primary concern is not whether it is management’s responsibility or auditors’ responsibility, but rather the devastating impact that fraudulent financial statements can have on all the involved parties. The spate of accounting scandals and corporate failures that occurred around 2001 has brought unprecedented attention to the importance of fraud detection and deterrence.

In an attempt to restore investor confidence and order in financial markets, following the multi-billion dollar accounting scandals at Enron and WorldCom, Congress passed the Public Company Accounting Reform and Investor Protection Act of 2002, otherwise
known as the Sarbanes-Oxley Act (SOX). SOX is said to be the most comprehensive and important corporate governance reform since the Securities Exchange Act of 1934 was enacted to regulate securities markets and increase transparency of financial reporting. As a result of this enactment, the corporate accounting world has been redefined. Complying with the requirements of SOX has introduced many changes in firms such as financial reporting has become tighter, internal controls have improved, and there is more transparency in the overall auditing process. In addition, the Act exposes both management and auditors to greater litigation risks and stricter penalties for misconduct.

Despite increased anti-fraud measures taken by the government to combat fraud (such as the enactment of SOX), financial statement fraud remains a public concern. Experts agree that absolute fraud prevention is an unrealistic and unobtainable goal. As a matter of fact, no one can guarantee that fraud will be either prevented or detected in a timely manner.

One of the most difficult tasks in detecting financial statement fraud is to identify symptoms of fraud because of the nature of financial statement fraud. Some factors will be present when no fraud exists. For example, GAAP violations need not involve fraud in situations where management believes that the application of GAAP is inappropriate for the company’s unique circumstances and discloses the deviation from GAAP in annual reports, and only a small number of symptoms may present themselves when fraud is occurring (for instance, fraud can occur without any manipulation when management deliberately omits known contingent liabilities or related party transactions from notes to the financial statements contained in annual reports and impact of this fraud is indirect and difficult to measure). Since symptoms can be caused by legitimate factors, the
presence of a symptom does not always mean that fraud exists. Fraud symptoms cannot
easily be ranked in order of importance, nor can they be combined into effective
predictive models. Their relative importance varies widely. Fraud detection is
complicated by the fact that there is little consensus among experts about symptoms that
consistently indicate fraudulent behavior. However, it has been widely acknowledged that
fraud symptoms often exhibit themselves through changes in financial statements.
Therefore, it is crucial to examine financial statements in order to observe these
symptoms of fraud.

The difficulty of fraud detection is further exacerbated by the fact that even when the
annual reports conform to U.S. GAAP and are within the confines of reasonable
interpretation of the GAAP rules, they can be misleading. This is due to the fact that the
U.S. accounting standards are “rules based” rather than “principles based.” Since rules do
not cover every conceivable situation, it is possible for the companies to be creative in
financial measurement as well as disclosures. For example, earnings manipulation is
different from earnings management. In case of earnings management, companies try to
smooth earnings over a period of time to maintain regularity in flow of earnings, as
investors perceive consistency of earnings favorably. However, in case of earnings
manipulation, companies try to hide unfavorable information by using deceptive and
improper mechanisms such as making up numbers and thus do not present a full and fair
representation of its financial performance. In the case of Enron, there was earnings
manipulation that added a billion dollars of false earnings to its books through bogus
transactions. Detecting fraud requires one to look beyond the numbers because financial
numbers can be easily camouflaged. Due to the relative infrequency of fraud and the
shortage of knowledge concerning the characteristics of management fraud, if fraud is being committed and external auditors are not vigilant (when using normal audit procedures and thus fail to detect fraud), it is difficult for an average investor to detect problems. Even when auditors are vigilant, if management decides to commit fraud, fraud can not be prevented because companies use creative methods to present annual reports in a way that conceals as much as they reveal (Griffiths, 1986).

1.2 PROBLEM STATEMENT

A major thrust of the earlier studies of fraud detection has been on the examination of financial factors such as financial ratios and metrics that are extracted from financial statements. In fact, Kaminski et al. (2004), have suggested a limited ability of financial variables to predict fraud. In addition, earlier studies on fraud detection have left out a key component of financial statements, qualitative textual content. Financial statements communicate quantitative information and qualitative narratives as well as forward-looking information in the company disclosures. In annual report narratives, explanations are often expressed within the framework of accounting language. In other words, information on fraud indicators is not explicitly stated in annual reports but rather hidden in such documents. Systematic and objective analysis of large volumes of text data in annual reports (also known as ‘Form 10-Ks’) is important because only a tiny fraction of all corporate information is quantitative in nature.

The basic premise of this study is that valuable information is buried in textual disclosures by the companies in documents such as ‘Form 10-Ks’ and ‘Form 8-Ks.’ Most of the research thus far, however, has focused on analyzing the numerical data such as
financial metrics and ratios (Green and Choi, 1997; Fanning and Cogger, 1998; Lee et al., 1999; Spathis, 2002; Spathis et al., 2002; Kaminski et al., 2004; Kirkos et al., 2007; Hoogs et al., 2007). Companies are, however, very innovative in camouflaging their financial statements and gerrymandering annual reports to hide the true picture. We believe that the textual documents released by companies contain indicators that the company may have carelessly left exposed or strategically placed phrases using selective accounting language. An examination of these cues, hidden in the qualitative content of annual reports, may reveal new, interesting, and useful information for fraud detection.

Prior research highlights the importance of textual portions of annual reports to prime users of financial accounting information such as investors and financial analysts (Abrahamson and Amir, 1996; Bryan, 1997; Rogers and Grant, 1997). These studies have utilized only portions of annual reports due to implementation constraints to predict outcomes such as bankruptcy, financial distress, company performance and so forth. For example, methodology of some of the earlier studies involved manual examination of the qualitative content of annual reports. In contrast, other studies used a hybrid of automated and manual tools to do qualitative analysis, but the researchers limited the analysis to only those portions of annual reports, which they suspected to be relevant to their studies to factor in for the manual component of their analysis. Very few studies address the annual report as a whole, in terms of the integration of the messages across the various parts of the report. With the exception of Cecchini (2005), none of the previous studies have utilized the qualitative content of annual reports to detect fraud.

The aim of this dissertation is to contribute to (1) the development of fraud detection methods to increase the possibility of fraud detection and (2) to enhance the timeliness of
fraud detection so that early warning signs can be detected. Early detection of fraud provides the best opportunity to mitigate the damage that can be caused (as noted by Persons, 1995; Fanning and Cogger, 1998; Lee et al., 1999; Kaminski et al., 2004; Hoogs et al., 2007).

This dissertation proposes a method to proactively detect fraud by examining the qualitative content of the annual reports as a whole. In the qualitative textual analysis, it examines both the verbal content and the presentation style of the annual reports and explores the linguistic features that distinguish fraudulent annual reports from non-fraudulent annual reports. It also creates a classification metric for levels (stages) of fraud for early prediction by examining changes in the qualitative content of annual reports for pre-fraud, fraud, and post-fraud periods of fraud companies.

Darrough and Stoughton (1990) used a game theoretic approach to show that companies have an incentive to engage in selective disclosures in order to meet their own objectives. Contrary to the theory of discretionary disclosure, which relates to new information, this study extends previous research on the theory of selective disclosure, where management uses creative accounting practices in disclosing past information. In the following section, we discuss the importance of this research.

1.3 IMPORTANCE OF THIS RESEARCH

Even though the incidence of financial statement fraud is low relative to the total number of companies with shares that are traded on U.S. exchanges, the economic losses associated with them are monumental (e.g., the demise of Enron and Arthur Anderson LLP). Wells (1997) estimated that fraud costs US business more than $400 billion
annually. AICPA (2005) reports that financial statement fraud causes a decrease in market value of stock of approximately 500 to 1,000 times the amount of the fraud. For example, a $7 million fraud can cause as much as $2 billion drop in the value of stock. Clearly, the consequences of financial statement fraud are enormous, even terminal in the case of some companies.

In spite of the fact that the intent behind SOX was investor protection from fraud and a guarantee of the accuracy of financial reports, numerous cases of fraudulent financial reporting have surfaced in the post-SOX period [e.g., the AIG scandal (Business Week, April 2005)]. From 2000 through 2006, the SEC issued 344 Accounting and Auditing Enforcement Releases (AAERs) relating to financial statement fraud as shown in Figure 1.1. It should be noted that several of these AAERs were issued to same company and not all investigations of these AAERs resulted in finding of financial statement fraud.

![Financial Statement Fraud AAERs Issued by the SEC, by Year of Issue](image)

**FIG. 1.1 - Financial Statement Fraud AAERs Issued by the SEC, by Year of Issue**

*Source: Deloitte Forensic Center (2007)*
A survey conducted by the Association of Certified Fraud Examiners (2004) pointed out that fraud is more pervasive today in the post-SOX period than it was in 2002 and there is every indication that it might get worse. The same survey revealed that the median fraud costs to US business are more than $600 billion per year.

Implementation of SOX has imposed a high burden of regulations and compliance in the corporate world. Unfortunately, all these complex and costly measures have been unable to stop companies from breaking the law. According to AMR Research (2006), SOX compliance is costing about 50 times more than was estimated in 2002, exceeding about $6 billion. Since SOX compliance is mandatory for public companies only, many companies have started deregistering from the SEC to avoid the obligation of providing information mandated by the securities laws to investors. Leuz et al. (2007) found that delisting of public companies tripled in 2003 from 2002 and approximately 200 U.S. companies deregistered their common stock for reasons other than a liquidation, merger, or acquisition in 2003 alone. It was noted that some companies delisted to save money and other companies delisted to avoid outside monitoring and public scrutiny (Engel et al., 2007; Leuz, 2007).

Experts agree that corporate governance mechanisms are necessary, but believe that overrelying on them is not going to diminish the possibility of another disaster, such as those that occurred at Enron and WorldCom. As a matter of fact, all companies involved in major corporate scandals had fairly elaborate corporate compliance programs in place. Rezaee and Jain (2006) doubt that the intent of SOX was to combat fraud and instead present evidence that this may be coincident with the tumble in stock values. They argue that if stock values were high, then nobody would care about issues like increased
disclosure and greater transparency. Donaldson (2005) cautioned that simply rearranging the chairs at the higher echelons of a company will not prevent the types of fraud that have occurred over the past several years. He further notes that it is overly optimistic to assume that corporate reform efforts such as SOX can fix everything so that the next round of scandals will not happen.

According to Lynn E. Turner, the former chief accountant of the SEC, financial statement fraud has cost investors more than $100 billion during the past several years, with the Enron scandal causing a loss of about $80 billion in market capitalization to investors, including employees and pensioners (The Washington Post, February 2002). As reported in Public Citizen (September 2002) some of the other biggest losses came to shareholders in Tyco ($84.2 billion), Lucent ($55.5 billion), WorldCom ($26.9 billion), Xerox ($9.8 billion) and Qwest ($9.8 billion). These studies and statistics provide only underestimated direct economic losses that resulted from financial statement fraud. The overall cost of fraud including direct economic losses resulting from it and indirect costs such as adverse impact on confidence of investors, employees’ morale, customers’ goodwill, and suppliers’ trust as well as increased legal costs, insurance costs, and loss of productivity is over double the amount of missing money and assets (Farrell and Healy, 2000). Next, we provide a brief overview of our research methodology.

1.4 BRIEF OVERVIEW OF RESEARCH METHODOLOGY

The methodology used in this study was implemented using Natural Language Processing tools. It deals with analyzing, understanding and generating the language. It includes syntactic, morphological, semantic, and phonological analyses. The application
of Natural Language Processing tools for fraud detection is a fertile research area that
should be investigated to the fullest possible extent. Unlike research in other well-
examined fields of accounting and finance, such as bankruptcy prediction (Zhang et al.,
1999; Lensberg et al., 2006), credit risk assessment (Henley and Hand, 1996; Fritz and
Hosemann, 2000; West, 2000; Doumpos and Pasiouras, 2005), detection of earnings
management (Peasnell et al., 2000), financial distress prediction (Anandarajan et al.,
2001), prediction of fraudulent insurance claims (Viaene et al., 2002), and prediction of
currency crises (Liu and Lindholm, 2006), research on detecting financial statement fraud
using machine learning based classifiers such as Naïve Bayes classifiers, neural
networks, or support vector machines is a relatively new phenomenon.

This dissertation presents a methodology to proactively detect fraud by examining
qualitative content of annual reports using Natural Language Processing (NLP) tools. The
methodology employs machine learning techniques to build an automated classifier that
can predict the likelihood of fraud. In this study, a fraud detection model is built using
Support Vector Machines (SVM), a supervised machine learning technique that
incorporates domain knowledge while the learning task is undertaken from a set of
positive and negative examples provided as the training set. This research seeks to
explore linguistic features that differentiate fraudulent annual reports from non-fraudulent
annual reports.

Once the learning is successful, SVM is able to successfully classify unlabelled annual
reports in the testing dataset as fraudulent or non-fraudulent. The correctness of fraud
predictions is then evaluated against the correct fraud classes of the testing dataset.
Several standard evaluation measures such as accuracy, precision, recall, and F-measures
are used (Manning and Schutze, 1999). These evaluation measures presuppose that each document (annual report) belongs to only a single class (fraudulent or non-fraudulent). The fraud classifier is also trained on pre-fraud and post-fraud data of fraudulent companies to detect early warning signs of fraud. The accuracy of its predictions is tested on a hold-out sample using the measures mentioned earlier.

1.5 OVERVIEW OF THE THESIS

This thesis is divided into eight chapters as outlined below.

Chapter 1 discusses the motivation, the research problem and importance of the problem along with the background information on fraud and a brief overview of the research methodology.

Chapter 2 reviews the fraud detection literature and summarizes key observations and results. It presents earlier fraud detection models, pointing out why those approaches are inadequate and recent phenomenon of studies employing Artificial Intelligence (AI) and machine learning techniques to model similar problems in the domains of accounting and finance. It also presents an overview of the body of literature that has examined qualitative content of company reports and other company disclosures using different NLP techniques.

Chapter 3 proposes a methodology to detect fraud from the qualitative content of company annual reports using NLP tools. First, it provides an overview of the fraud classification problem and proposes a fraud detection model. Second, it evaluates various classifiers and discusses why SVM has been chosen to develop this model. Third, it
discusses how input features will be selected and leads into the choice of variables and hypothesis development. It then describes the procedure of hypothesis testing that we employed. Fourth, it describes the sample, data collection, and pre-processing of the dataset. Fifth, it presents a baseline approach, preliminary baseline results and how the classifier will be trained. Sixth, it discusses the measures to test the prediction accuracy of the fraud detection model. Finally, it presents some concluding remarks.

**Chapter 4** discusses the sample selection process and data collection methodology in detail. It first explains the data definitions and rationale behind them. It also describes the sample that was finally selected for building the fraud detection model.

**Chapter 5** discusses the baseline results. It first explains the baseline approach that was used to run baseline experiments. After discussing the baseline features, it presents the baseline experiments that were run with the Naïve Bayes classifier for detecting fraud and for detecting early symptoms of fraud. It then presents the baseline results when these features were fed into an SVM classifier, the main classifier used in this study. The purpose of conducting the set of experiments in this chapter is to show to what extent it is possible to detect fraud when no feature extraction or feature selection is performed.

**Chapter 6** discusses the features and the results of the classifier experiments in detail. It first discusses the features that are used to examine the presentation style and content of the annual reports for detecting fraud and stages of fraud. In addition, it also describes the feature extraction process and the results of the classifier experiments that we obtained with these features. After this, it provides an overview of the feature selection process.
and finally presents the subsets of features that were found to be most useful for recognizing fraudulent and non-fraudulent annual reports and different stages of fraud.

Chapter 7 gives the results of testing the hypotheses. It also discusses the results, explaining whether or not the hypotheses can be supported.

Chapter 8 provides a brief overview of main findings, implications, and potential future work. First, it provides a summary of the research and summarizes the key results. Second, it discusses the implications of the findings of this research for both academicians and practitioners and outlines future research that builds on the current work presented in this thesis. Third, it presents some inherent limitations. Finally, it offers concluding remarks.
CHAPTER 2
LITERATURE REVIEW

This chapter reviews the fraud detection literature and summarizes key observations and results. Section 2.1 presents early fraud detection models and explains why those approaches are inadequate. It then describes more recent studies employing Artificial Intelligence (AI) and machine learning techniques to model similar problems in the domains of accounting and finance. Section 2.2 presents a critical overview of the literature in the analysis of qualitative content of company reports and disclosures using various NLP techniques. Section 2.3 compares our work in this dissertation to related works in the area of fraud detection and highlights important differences. Section 2.4 summarizes our key findings.

2.1 APPROACHES TO FRAUD DETECTION USING AI AND NON-AI TECHNIQUES

Until recently, most models of fraud detection were based on variations of logistic regression, linear discriminant analysis, and probit analysis. However recently the trend has been to model fraud detection using artificial intelligence (AI), data mining, and machine learning techniques. One reason for this shift is attributable to limitations of some of these earlier models. For example, Hoogs et al. (2007) have noted that discriminant analysis does not handle time series data well and that better models are needed to include year-to-year changes to capture progression of fraud, bankruptcy, and
financial distress. In the case of logistic regression models, Hoogs et al. (2007) also noted that, these models are impractical for applications with high dimensional datasets. Compounding these problems is the issue of the limited ability of these models to handle missing values in training data often encountered in publicly available financial data.

Another reason for this shift may be to expand applications of AI technologies in the domain of accounting, particularly auditing and assurance. In the remainder of this section, we present prior studies that have attempted to build fraud detection models using non-AI and AI techniques, in tandem with their results.

Uecker et al. (1981) conducted an experimental study to investigate whether greater responsibility being thrust upon both internal and external auditors for the prevention and detection of corporate fraud would function as a deterrent for management to commit fraud. Their results suggest that managers contemplating acts of corporate fraud are not deterred by the presence of internal and external auditors, and an increase in the perceived “aggressiveness” of the internal and external auditors by management did not increase or decrease the occurrence of corporate irregularities. The findings of this study have important implications for the post-SOX period in ascertaining the potential effectiveness of the greater accountability imposed on both management and the auditing profession in the prevention and detection of corporate fraud.

Matsumura and Tucker (1992) carried out game-theoretic analyses and economic experimentation to understand the interaction between a manager who can commit fraud and an auditor who can test for fraud by focusing on the effects of economic incentives on fraud commission and detection. In the game, the manager chose the probability of
committing fraud whereas the auditor decided to perform tests of transactions. They considered four independent variables—the auditors’ penalty, auditing standards requirements, the quality of the internal control structure, and the audit fees—to assess their effects on tests of transactions (to detect irregularities), fraud detection, and incidence of fraud. They found that increasing the auditors’ penalty decreased fraud and increased fraud detection and so did the strength of the internal control; strong internal controls resulted in increased fraud detection. In addition, their study found that managers committed fraud less frequently in the presence of strong internal controls.

Persons (1995) developed a stepwise-logistic fraud detection model using quantitative financial metrics such as financial leverage, capital turnover, and asset composition. He used a matched pair sample of 103 firms for the fraud year and 100 firms for the year preceding the fraud year, as identified by SEC enforcement actions, and an equal number of no-fraud firms matched on industry and time period. The fraud year model correctly classified 47% of the fraud firms, while misclassifying 14% of the no-fraud firms. The preceding year model correctly classified 64% of the fraud firms, while misclassifying 21% of the no-fraud firms. He concluded that, in addition to financial metrics used in the study, firm size also played an important role in fraud detection.

Beasley (1996) built a logit cross-sectional regression fraud detection model to examine the relationship between board of director composition and the occurrence of financial statement fraud. Using a matched-pair sample of 75 fraud and 75 no-fraud firms matched on size, industry, national exchange (where common stock was traded), and time period, he found that the proportion of outsiders on the board of directors was lower for firms experiencing financial statement fraud even after controlling for the presence of an audit
committee. In addition, factors such as board size and certain outside director characteristics were also found to affect the likelihood of financial statement fraud. He concluded that outside director characteristics (such as differences in personality, management style, and other behavioral characteristics), need to be further explored to corroborate his findings.

Hansen et al. (1996) developed a generalized qualitative-response model using probit and logit techniques to predict the presence of management fraud. They used a set of data that was developed by an international public accounting firm, which consisted of 77 fraud cases and 305 no-fraud cases. Their results indicate good predictive capability for both symmetric and asymmetric misclassification cost assumptions.

Kwon and Feroz (1996) compared the results of a multilayered perceptron neuron network fraud detection model with a logit model trained on a matched-pair sample of 35 fraud firms identified in SEC AAERs and 35 control firms matched on industry, size, and time period. The fraud model based on a neural network correctly classified 88% of the fraud firms on average, whereas the logit analysis model correctly classified only 47% of the fraud firms on average. They concluded that a multilayered perceptron neural network model was more powerful for detecting problems with statistical uncertainty. They further noted that non-financial information (such as frequency of turnovers in chief financial officers, chief executive officers, and auditors of a firm) increased the predictive ability of the model compared with the use of financial (such as profitability, sensitivity, difficult to audit, and going-concern) ratios alone.
Deshmukh et al. (1997) developed a fuzzy reasoning system based on the research on red flags for assessing the risk of management fraud. The fuzzy reasoning system employed ‘Attitude,’ ‘Condition,’ and ‘Motivation’ as the main fuzzy parameters and the variable ‘Management Fraud’ as the output variable. The possible outcomes for the ‘Management Fraud’ variable were ‘High,’ ‘Medium,’ and ‘Low.’ The sample consisted of total of 382 cases, divided as 254 low-risk cases, 38 moderate-risk cases, and 90 high-risk cases. In the case of low risk, the fuzzy reasoning system identified 253 low-risk cases (99.6% hit ratio), 0 moderate-risk cases (0%), and 9 high-risk cases (10%). In the case of moderate risk, the fuzzy system identified 35 low-risk cases (92%) and only 3 high-risk cases (8%). In the case of high risk, the fuzzy system identified 57 low-risk cases (63%) and 24 moderate-risk cases (27%). They observed that the fuzzy system tended to underestimate the risk in case of moderate-and high-risk areas.

Eining et al. (1997) designed a fraud risk assessment model using an expert system to aid auditors in the detection of fraud. They argued that most auditors have never encountered management fraud before and, as a result of this lack of prior experience, have limited ability to detect fraud of this nature. Their results showed that use of an expert system helped auditors better discriminate in situations with varying levels of management fraud-risk and take appropriate audit actions.

Green and Choi (1997) developed a neural network fraud detection model employing endogenous financial data that consisted of five financial ratios (Allowance for doubtful accounts/Net sales, Allowance for doubtful accounts/Accounts receivable, Net sales/Accounts receivable, Gross margin/Net sales, Accounts receivable/Total assets) and three accounts (Net sales, Accounts receivable, Allowance for doubtful accounts) as
input. Their sample consisted of 86 fraud firms accused of management fraud in SEC AAERs and 86 no-fraud firms that were matched on the basis of year, size, and industry. The learning samples used to train the model were relatively small thus limiting the external validity of the study. Furthermore, neural networks suffer from an inherent weakness of lacking explanatory capabilities and thus justification for the results could not be obtained. They concluded that neural networks have good capabilities when used as a fraud detection tool and their reliability can be further improved by incorporating qualitative factors as input variables.

Fanning and Cogger (1998) compared the results of a fraud detection model using an artificial neural network approach with those from stepwise logistic regression, linear discriminant analysis, and quadratic discriminant analysis. They used a matched-pair sample of 102 fraud companies identified in SEC AAERs and 102 no-fraud companies controlled for industry, fiscal year end and company size. The input vector to the neural network model consisted of financial ratios such as accounts receivable to sales, net property, plant and equipment to total assets, debt to equity, and other suggested red flags for fraudulent financial statements such as percentage of outside directors, geometric growth rate and having a non-Big Six auditor. Their neural network model demonstrated superior ability in comparison to standard statistical methods in detecting fraud and accurately classified 69% of the fraud companies while misclassifying 20% of the no-fraud companies in their training data, and accurately classified 66% of the fraud companies, while misclassifying 41% of the no-fraud companies in their testing data.

Summers and Sweeney (1998) used a cascaded logit model to examine the relationship between insider trading and fraud, and classified firms into fraud and no-fraud categories
based upon financial statement characteristics such as financial condition, financial performance, growth, receivables/inventory, auditor change and insider trading variables such as the dollar amount, number of shares, and number of transactions for purchasing and selling by insiders during the year of fraud occurrence. They used a matched sample of 51 fraud companies found in the *Wall Street Journal* and 51 no-fraud companies matched on industry and size to build their model. They found that, in the presence of fraud, insiders in companies with fraudulent financial statements strategically reduce their net position in the company’s stock by selling more stock. However, the other predicted relationship between significant decline in insiders purchasing activity of company stock during the period of fraud’s occurrence was not supported by their empirical findings.

Lee et al. (1999) used a logistic regression fraud detection model to examine if the difference between earnings and operating cash flow is an indicator of fraud in the year prior to the discovery of fraud. They used a peer sample approach in which they included 56 target companies identified in SEC AAERs and the Wall Street Journal Index and 564 no-fraud firms matched on industry and time period. Their model consisted of quantitative variables such as earnings, earnings minus cash flow, leverage, sales growth, market return, retained earnings, market value, and indicator variables for auditor change, etc. Their model accurately classified 73% of the fraud firms, while misclassifying 10% of the no-fraud firms at the 10% probability cut-off, and accurately classified 61% of the fraud firms, while misclassifying 4% of the no-fraud firms at the 20% probability cut-off.

The Statement on Auditing Standards (SAS) No. 82, issued by the Auditing Standards Board in 1997, regarding consideration of fraud in a financial statement audit, requires
auditors to assess the risk of fraud on every audit and consider both the internal control system and management’s attitude toward controls when making this assessment. In light of this pronouncement, Caplan (1999) developed a model in which he explains that managers can commit fraud regardless of control strength. However, managers might choose weak internal controls in order to “hide” fraud, and the use of control strength as a “red flag” in fraud risk assessments can be a potential indicator of fraud. He concluded that auditors should observe both the actual strength of controls and the manager’s choice of controls. The probability that an auditor will detect fraud is conditional on the existence of fraud and the auditor’s effort, and audit risk in the model comes from failing to detect fraud (type II errors).

Abbot et al. (2000) conducted a statistical regression analysis to examine if the existence of an independent audit committee (measured in terms of its independence and activity) mitigates the likelihood of fraud. They used a sample of 78 firms identified in SEC AAERs matched with 78 non-sanctioned firms that are similar in size, industry, national exchange and time period. They concluded that companies with audit committees that consist of independent members and meet at least two times in a year are less likely to be sanctioned for fraudulent financial reporting.

Bell and Carcello (2000) developed a fraud detection model using logistic regression that was conditioned on the presence or absence of various fraud-risk factors. They used a sample of 77 fraudulent engagements and 305 non-fraudulent engagements. Some of the fraud-risk factors included in their final model were a weak internal control environment, rapid company growth, aggressive management attitude towards financial reporting, evasive approach of management towards auditors, undue emphasis by management on
meeting earnings projections, and inadequate or inconsistent relative profitability. Their results showed that, in case of fraudulent observations, the logistic model was more accurate than practicing auditors in assessment of risk. However, there was not much difference between the model assessments and those of auditors in the case of non-fraudulent observations.

Motivated by concerns regarding auditors’ role in fraud detection, Braun (2000) investigated the effects of time pressure on auditors’ attention to indicators of potential fraudulent financial reporting in a dual cognitive task environment. He predicted that, under time pressure, auditors’ attention would become focused on the dominant task at hand at the expense of attention to the subsidiary task and thus auditors would be preoccupied with the task of accumulating documentary evidence regarding the frequency of misstatements. In the process, the qualitative aspects of misstatements indicative of potential fraudulent financial reporting would be overlooked. The results of the study were consistent with his predictions and he found that the accuracy of documenting the evidence on frequency of misstatements did not suffer under time pressure but auditors’ professional skepticism did decline. However, this study did not consider the effects of experience, training, and the nature of risk factors on auditors’ ability to detect fraudulent financial reporting under time pressure.

Nieschwietz et al. (2000) conducted empirical research on external auditors’ ability to detect financial statement fraud. In this study, they synthesized various studies that provide evidence as to how well auditors use the methods prescribed by the profession for assessing fraud risk. They also identified empirical predictors of fraud and investigated auditors’ unaided use of fraud cues to assess fraud risk as well as auditors’
reliance on technically-aided fraud-risk assessment models. Their findings suggested that more research is required that focuses on increasing the auditors’ reliance on fraud detection aids. In addition, they pointed out that research is needed on mechanical and cognitive models that assist auditors in fraud-risk assessment and link risk assessments to audit plans. Finally, the authors emphasized the need to improve the effectiveness of audit plans to detect fraud and suggested that behavioral research within a game-theoretic context will help accomplish this task.

Albrecht et al. (2001) provided empirical evidence regarding the plausibility of using fraud hypothesis testing as a method of detecting fraud that involved identifying ‘red flags’ or indicators of fraud and following up on them to determine whether they represent fraud or are the result of other non-fraud factors. This approach required the formulation and testing of many null hypotheses, each stating that there is no fraud of a particular type. The fraud hypothesis testing approach required the identification of fraud that may exist in particular situations, the identification of red flags that each of those particular instances of fraud would create, and the design of customized queries to search for those specific red flags or combinations of red flags to test each of the specific fraud hypotheses. This study differed from previous fraud research in that it attempted to formalize the use of red flags to proactively detect fraud that has not yet been discovered.

Johnson et al. (2001) designed a computational model of financial fraud detection that employs the heuristics gained from experience with deceptions in everyday life. They defined the financial deception tactics used by management in the domain of accounting such as masking, dazzling, decoying, repackaging, mimicking, and double playing and, to counter each of these deception tactics, they modeled detection tactics in a ‘competence’
model. The input to this model consisted of 24 data items from a set of financial statements; its diagnostic output, audit opinion (‘misleading,’ ‘unqualified+,’ ‘unqualified’), was an indication of the extent to which the financial statements are a fair representation of the company’s financial conditions. For instance, a ‘misleading’ audit opinion indicated that the auditor rendered either an ‘adverse’ or ‘Non-GAAP’ opinion, whereas an ‘unqualified+’ opinion indicated that the auditor felt that the case was fairly represented, yet there are areas of concern that need to be explicitly highlighted. The model was progressively modified by introducing failures to detect deception and applying heuristics to the specific context of financial statement fraud until it could successfully detect fraud in the six experimental cases (four fraud and two no-fraud) given to auditors.

Spathis (2002) developed a fraud detection model using stepwise logistic regression. The sample consisted of 38 Greek manufacturing firms with fraudulent financial statements and 38 no-fraud firms. The input vector to the model consisted of financial ratios such as the inventories to sales ratio, the total debt to total assets ratio, the working capital to total assets ratio, the net profit to total assets ratio, and financial distress (Z-score), which were derived from the published financial statements. The results indicated that companies with high inventories with respect to sales, high debt to total assets, low net profit to total assets, low working capital to total assets and low Z scores were more likely to falsify financial statements. The model correctly classified the total sample with accuracy rates exceeding 84%. However, he did not use a hold-out sample to validate the prediction accuracy of the model. He concluded that designing a more powerful analytical tool with
larger samples for detection of fraudulent financial statements would be useful for validating these results and increasing the prediction accuracy.

Viaene et al. (2002) examined several AI techniques to detect fraudulent insurance claims. Their results indicated that the non-linear techniques such as neural networks did not perform as well as linear techniques. The poor performance of non-linear techniques was attributed to a lack of domain specific data in the limited test scenario. They concluded that if domain specific knowledge is incorporated in these models, then non-linear techniques provide more flexibility in developing fraud classification models.

Lin et al. (2003) employed a hybrid class of intelligent systems called fuzzy neural networks (FNN) that integrate fuzzy logic with artificial neural networks for detecting fraudulent financial reporting. Their sample consisted of 40 fraudulent firms and 160 non-fraudulent firms. They divided the sample evenly into a training set and a testing set, with each set consisting of 20 fraudulent and 80 non-fraudulent firms. The training set was used to develop a FNN and its performance was compared with that of a logit model using the testing set. They found that the FNN outperformed most statistical models and prior artificial neural network models. In the test set, FNN correctly classified 7 out of the 20 fraud cases with a hit ratio of 35 percent, and 69 out of the 80 non-fraud cases with a hit ratio of 86.3 percent. Its average accuracy was 76 percent. On the other hand, the logit model correctly classified only 1 out of the 20 fraud cases with a hit ratio of 5 percent, and 78 out of the 80 non-fraud cases with a hit ratio of 97.5 percent. The overall prediction accuracy of the logit model was 79 percent. They concluded that, for the same dataset, both the logit model and the FNN were successful in identifying non-fraud firms but, for fraud cases, the FNN achieved higher accuracy than the logit model.
Cullinan (2004) reviewed cases of financial statement misstatements that occurred at Enron, WorldCom, Sunbeam, Cendant, and Waste Management and presented a model of the auditors’ misstatement detection and reporting process, which broke down at multiple points among these cases. Each case represented a unique combination of an auditor failure to see the problem transactions, to recognize the transactions as misleading, and/or disclose the problem in the audit report. The paper concluded that most of the audit-related provisions of the Sarbanes-Oxley Act (SOX) are concerned with strengthening auditor independence, which was responsible for the Enron failure; however, there may be other aspects of the audit process that are associated with other recent frauds that need more meaningful legislative reform. For instance, he pointed out that SOX did not contain provisions designed to enhance the intellectual ability and diligence of auditors to recognize problems.

Kaminski et al. (2004) conducted an exploratory study to determine the fraud detection capabilities of ratio analysis. They compared a multitude of financial ratios to see if the financial ratios of fraudulent companies differ from those of non-fraudulent companies. This study examined an extended time period of both pre-fraud and post-fraud years. Statistically, there was not much difference in the ratios of fraud versus no-fraud firms. Those ratios found significant were not consistent across the time periods. A discriminant prediction model misclassified fraud firms from 58 percent to 98 percent of the time. Their results provide empirical evidence of the limited ability of financial ratios to detect and/or predict fraudulent financial reporting. However, the ratios selected for inclusion in this study were based on scattered heterogeneous empirical evidence and logical
inferences of accounts most likely involved in fraudulent financial reporting. Different results might have occurred if different ratios were selected.

Farber (2005) studied the linkage between the credibility of the financial reporting system and the quality of corporate governance mechanisms by investigating changes in these mechanisms subsequent to fraud detection. Consistent with prior research, his results indicate that fraud firms have poor governance relative to a control sample of no-fraud firms in the year prior to fraud detection. Specifically, fraud firms have fewer numbers and percentages of outside board members, fewer audit committee meetings, fewer financial experts on the audit committee, a small percentage of ‘Big 4’ auditing firms, and a higher percentage of CEOs who are also chairmen of the board of directors. His results indicated that fraud firms took actions to improve their corporate governance and, three years after fraud detection, these firms had governance characteristics similar to control firms but credibility was still a problem for these firms.

Rezaee (2005) shed light on the factors that might increase the likelihood of financial statement fraud. In light of the difficulties and costs associated with deterring financial statement fraud, he emphasized the importance of understanding five interactive factors [Cooks, Recipes, Incentives, Monitoring and End-Results (CRIME)] that can influence fraud occurrence, detection and prevention. He identified the top management team of publicly traded companies as the ‘Cooks’ of financial statement fraud. He explained that ‘Recipes’ of financial statement fraud could range from overstating revenues and assets to understating liabilities and expenses. Under ‘Incentives,’ he discussed the motivations for companies and their cooks to perpetrate financial statement fraud. Under ‘Monitoring,’ he demonstrated that lack of an effective corporate governance mechanism,
lack of oversight by an audit committee, and inadequate and ineffective internal control structures can contribute to financial statement fraud. For ‘End-Results’ he suggested that fraud vulnerability reviews and gamesmanship reviews should be performed both periodically and on an ongoing basis. He analyzed a sample of high profile alleged financial statement fraud cases to explain and justify these five interactive factors and stressed that ‘cooking the books’ is a crime that should have serious consequences.

Hoogs et al. (2007) used a genetic algorithm approach to detect financial statement fraud. Their sample included a target class of 51 companies accused by the SEC in AAERs and a peer class of 339 companies matched on industry and size. The input vector included 76 financial metrics and ratios. Their model accurately classified 63% of the target class companies and 95% of the peer class companies. They concluded that genetic algorithms provide a successful technique for detecting discriminatory patterns in challenging domains, and inclusion of qualitative indicators may provide additional discriminatory power.

Kirkos et al. (2007) explored the effectiveness of data mining classification techniques in detecting financial statement fraud. They compared the usefulness and performance of decision trees, neural networks and Bayesian Belief Networks (BBN) in detecting fraudulent financial statements. Their data consisted of 38 Greek manufacturing firms, which showed proof of involvement in issuing fraudulent financial statement, and 38 no-fraud firms. The input vector initially consisted of 27 financial ratios derived from financial statements. They reduced the dimensionality of the data by discarding variables that were found to be non-informative and 10 variables were chosen in the final feature set. Their results showed that, in terms of performance, BBN (90.3%) achieved the best
performance compared to neural networks (80%) and decision tree (73.6%) models. They concluded that enriching the input vector with qualitative information such as the composition of board of directors and previous auditors’ qualifications might improve the accuracy rate.

The literature shows that most fraud detection models used financial metrics and ratios. Some of these models also included event and company characteristic data, such as board of director composition, auditor changes, and management changes in addition to financial data. However, many of the results indicated high rates of false positives and false negatives. The limitations of these models to correctly predict fraud can have serious implications. For instance, false positives will discourage investors and lenders and will mean a lost opportunity for the company. In contrast, false negatives will not only adversely affect investors and lenders, but also increase the litigation risk for auditors. Typically, the cost of misclassifying a fraud company (i.e., a false negative) is higher than the cost of misclassifying a no-fraud company (i.e., a false positive), but it may depend on the consumers of a particular fraud detection model. For example, if the fraud detection model is to be used for making unalterable investment or lending decisions, then models are needed that can achieve high accuracy while keeping false positives at a minimum.

### 2.2 QUALITATIVE ANALYSIS OF COMPANY REPORTS

This section presents an overview of the body of literature that has examined the qualitative content of company reports and other company disclosures using different
techniques for applications in the domains of accounting, auditing and assurance. This stream of research can be broadly divided into three categories:

- Qualitative analysis using computerized text-processing tools
- Manual examination of qualitative content
- Use of a hybrid of automated and manual tools to do qualitative analysis

Singhvi (1968) examined the annual reports of U.S. and Indian companies to determine the adequacy of corporate disclosures from the perspective of investors. The dataset consisted of 155 U.S. companies and 45 Indian companies. She used an index of disclosure to evaluate the annual reports of these 200 companies. The index included 34 items of information, which were considered important for making rational investment decisions by financial analysts. Her findings indicated that corporations disclosing inadequate information are likely to be small in size and less profitable. She showed empirically that inadequate disclosure of information in annual reports is related to greater fluctuations in the price of the securities of these companies. She also noted that corporate management is less inclined to take the initiative in disclosing adequate information in annual reports to stockholders, if disclosure is left to its discretion. In this context, she also examined the reasons advanced by corporations for nondisclosure. Even though the most frequently advanced reason by management was that full disclosure results in a competitive disadvantage, she empirically showed that more profitable companies frequently disclose more information than less profitable companies. Finally, this study found that the disclosure of information in annual reports by the listed
companies in India is less adequate and less investor-oriented as compared with the companies listed on US exchanges.

Ingram and Frazier (1980) examined the content of firms’ environmental disclosures to determine if they related to the firms’ environmental performance. They compared the measured content to a set of performance indices derived from actual measurements of specific types of environmental activities and found that the discretionary section of the annual reports of poorer performers contained more narrative environmental disclosures than the discretionary section of the annual report of the better performers except for disclosure of litigation. Interestingly, litigation is the only category of environmental disclosure that is contained in audited financial statements, which may, in part, explain the reluctance of investors to use information contained in voluntary social disclosures in investment decisions.

Frazier et al. (1984) evaluated narrative elements in accounting reports using a system called “WORDS,” a content analysis system that was originally developed for use in a psychoanalytic environment. This system is based on word-frequency contiguity logic and derives statistical relationships between and among narrative words, which are descriptive of central themes in the text. Other accounting applications of ‘WORDS’ range from the work of Harway and Iker (1974) to that of Tennyson et al. (1990), which examined the relationship between narrative disclosures and bankruptcy. Despite its successful implementation, “WORDS” had certain limitations, which limited its shelf life. It had a 215-word limit for the maximum size of the intercorrelation matrix. The issue with its algorithm was that, after parsing, it added word frequencies and chose up to 215 words for statistical analysis, but words with high frequencies may not always be
those of greatest interest. Subsequently, other routines and screening techniques were added to extend its utility but empirical results of prior studies were disappointing, attributable to the method used.

In contrast to the studies where the authors did not look at cases in which firms were approaching bankruptcy, D’Aveni and MacMillan (1990) content analyzed shareholder letters to manually examine the differences in the focus of attention of top managers in surviving and failing firms. Their findings indicate that successful firms focus on external critical success factors such as customer needs and demand growth whereas failing firms deny crises and focus on internal short-term factors. In addition to these studies, some of the other key studies involving manual examination of the qualitative data included analysis of self-serving attributions in shareholders letters (studied by Clapham and Schwenk, 1991) and examination of data in the “Management Discussion and Analysis” section to predict future performance of companies (studied by Pava and Epstein, 1993).

Abrahamson and Park (1994) explored the relationship between casual attributions made by management in their letters to shareholders and company performance using a computer-assisted content analysis technique of over 1,000 presidents’ letters contained in annual reports. They found that outside directors and top corporate officers intentionally conceal negative organizational outcomes from shareholders. They found that this low disclosure is associated with subsequent selling of stock. A number of large-sample studies of letters to shareholders from company presidents have revealed similar results, that management attributes negative organizational outcomes to uncontrollable environmental causes in order to shift the blame for those outcomes away from themselves and attributes positive organizational outcomes to their own actions.
(Bowman, 1984; Aerts, 1994). Abrahamson and Amir (1996) also examined the qualitative content of letters to shareholders to identify high frequency words with negative connotations. However, it should be noted that these studies (Abrahamson and Park, 1994; Abrahamson and Amir, 1996) used a hybrid of automated and manual tools to do the qualitative analysis.

Boo and Simnett (2002) investigated the information content of management’s prospective comments (MPCs) disclosed in the annual reports of 140 Australian financially distressed companies to predict companies’ future viability. They contextualized the data and categorized MPCs as optimistic, pessimistic, mixed, and no MPCs. They found that financially distressed companies avoid disclosing MPCs in absence of an optimistic outlook and were significantly more likely to fail than companies that disclosed optimistic MPCs, and were as likely to fail as companies that disclosed pessimistic or mixed MPCs. Their results were consistent with other studies (Steele, 1982; Clarkson et al., 1994; Bryan, 1997) that investigated MPCs in concluding that such comments are informative with respect to the future performance of the company.

Kloptchenko et al. (2002a) combined data and text mining methods for analyzing financial reports to see if the textual part of the report contains richer information than the financial ratios to predict future financial performance. They used self-organizing maps to do quantitative analysis and, for qualitative analysis, they performed prototype-matching text clustering. The dataset consisted of quarterly financial reports of three leading companies (Nokia, Ericsson, and Motorola) in the telecommunications sector for the years 2000-2001. Their study showed that clusters from qualitative and quantitative
analysis do not coincide. They explained that these dissimilarities are due to the fact that the quantitative portion of annual/quarterly reports tends to state information about company’s past performance whereas the qualitative portion of these reports contains indications about future company performance, which can be inferred by examining the linguistic structure and written style of the text. They also examined the relationship between the written style of a report and facts stated by the numbers. They found that the tone of a written report tends to change before the actual financial numbers change. The tone tends to be closer to the company’s performance of the next period. They noticed that if the company’s performance will be worse during the next period, then the report of the current period gets more pessimistic, even though the actual financial performance remains the same. They emphasized that it is important to analyze text from these reports since written style of a financial report changes before a dramatic change occurs in company financial performance. The strongest limitation of their study was the small size of data, which they planned to overcome by expanding the study to a larger text collection. In another related study, Kloptchenko et al. (2002b) found that this methodology was incapable of identifying the important patterns for text content-based retrieval from scientific text corpora.

Rezaee et al. (2003) conducted a content analysis on audit disclosures of Fortune 100 companies to examine the quality of the report of the audit committee. The report of the audit committee is intended to ensure that financial statements are legitimate, the audit was thorough, and the auditors have no flagrant conflicts of interest that may jeopardize their objectivity, integrity, and independence. Their findings concluded that all companies examined with respect to audit committee composition, structure, meetings
and qualifications were in compliance with the requirements of SEC. The authors expected that more effective audit committee disclosures, in conformity with the provisions of the Sarbanes-Oxley Act of 2002, would improve the trust and confidence in corporate governance, the financial reporting process and audit functions.

Cecchini (2005) used support vector machines to examine the risk of financial events such as management fraud, bankruptcy, and restatements. For detecting management fraud, he used a dataset of 122 companies, which consisted of 61 fraudulent and 61 non-fraudulent companies. The fraudulent companies were gathered using SEC AAERs and Accounting and Auditing Association Monograph. The fraudulent companies were matched with non-fraudulent companies based on SIC code, year, and total assets values. The experiments were run using Financial Kernel, Text Kernel, and Combination Kernel. The input vector of the fraud detection model included both quantitative financial variables (account ratios) and qualitative text variables (concept vector which was preprocessed using accounting ontology) for fraudulent and non-fraudulent public companies. The financial component of the input vector consisted of ratios and year-to-year changes of the ratios for two years for attributes such as sales, accounts receivable, inventory, total assets, and capital expenditures. The text analysis was limited to the ‘Management’s Discussion and Analysis of Financial Condition and Results of Operations’ (MDNA) section of the annual reports. The text vector represented the number of times each ontology concept such as ‘calculate,’ ‘security,’ ‘measure,’ ‘journal,’ ‘section,’ and ‘liquidation’ was encountered in the text of a company’s MDNA section. The results achieved by the Financial Kernel for fraud detection were very strong (95.9% 10-fold cross validation accuracy), whereas results achieved by the Text Kernel
for 10-fold cross validation were only 54.39%. The results achieved by the Combination Kernel were also strong due to the Financial Kernel component. However, the results achieved by the Combination Kernel were lower than the results achieved by the Financial Kernel alone.

Some behavioral studies have been conducted to understand how behaviors are expressed in texts and if there are any relationships between these behavioral expressions and actions. Keila and Skillicorn (2005) found that deceptive writing is characterized by a reduced frequency of first-person pronouns and a high frequency of negative emotion words and action verbs. Using deception theory, they applied a deception model to the Enron email dataset, and then applied singular value decomposition to elicit correlation structure between emails. Stolfo et al. (2006) used a combination of behavior models of user email accounts to detect early onset of a viral propagation without “content-based” or “signature-based” analysis that is commonly used in virus scanners.

### 2.3 COMPARISON WITH RELATED WORK

The major thrust of prior studies on fraud detection has been on quantitative analysis using financial metrics and ratios. Some of the earlier studies on fraud detection have also focused on empirically 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. 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). Kaminski et al. (2004) also concluded that these conventional quantitative financial factors are
inadequate for predicting fraud. In contrast, we use the verbal, qualitative (non-quantitative) content of the annual reports to build fraud detection model, which performed better than earlier fraud detection models (see Chapter 6 for more information on classifier results). Moreover, some of these earlier fraud detection models were based on small sample sizes.

Even though, researchers have examined the qualitative portion of annual reports to predict outcomes such as bankruptcy and future company performance, very few studies have used the entire qualitative content of annual reports to predict these outcomes. These studies have utilized only portions of annual reports due to implementation constraints. The methodology of most of the earlier text analytical studies depended on manual collection or manual analysis or use of a hybrid of manual and automated tools to examine qualitative content. Manual examination of qualitative content can be very tedious, time-consuming and expensive. In contrast, our research takes advantage of the advancements in artificial intelligence field to examine the entire textual content of annual reports.

Cecchini (2005) has also used support vector machines to examine the risk of financial events including management fraud. However, our research is different from Cecchini’s work in many respects. First, Cecchini’s qualitative analysis for fraud detection was limited to only the ‘Management’s Discussion and Analysis of Financial Condition and Results of Operations’ (MDNA) section of annual reports. In contrast, our research uses the entire qualitative content of annual reports to discover complicated patterns indicative of fraud. It is difficult to pinpoint a particular section of the report where fraud symptoms will occur. Hence, we examined the entire qualitative content, as we did not want to
throw away any useful content. Our goal was not to just extract useful information from text but also to identify hidden indicators of fraud. Second, Cecchini’s textual analysis involved examination of the verbal content. In contrast, we not only examined the verbal content (where we looked at the content words, frequencies of usage, word patterns, etc.) but also the presentation style of the annual reports. Third, we conducted a linguistic analysis to explore linguistic features [such as voice (active versus passive), uncertainty markers, readability index, tone, usage of proper nouns, type-token ratio, etc.] that could distinguish fraudulent annual reports from non-fraudulent reports whereas Cecchini simply examined the frequency of concept words in the MDNA section of annual reports. Our initial baseline results with a “bag of words” approach were similar to Cecchini’s results (54.39%), which were not good, but when we incorporated linguistically motivated features, we were able to improve the prediction accuracy of our fraud detection model from baseline performance (see Chapter 6 for results). Fourth, we used a peer set for each fraud company as opposed to a matched-pair dataset used by Cecchini. Fifth, our dataset was larger and it included 126 fraudulent companies and 622 non-fraudulent companies as opposed to 61 fraudulent and 61 non-fraudulent companies used by Cecchini.

2.4 SUMMARY

The major thrust of prior studies has been on empirically examining the relationship between fraudulent financial reporting and quantitative indicators such as composition of boards of directors, insider trading, auditor rotation, or financial restatements. Even though researchers have examined qualitative components of annual reports, very few studies have used the entire qualitative content to predict fraud. Building on prior studies,
the present study contributes to the literature gap by using an alternative methodology for examining and detecting fraud.
CHAPTER 3
METHODOLOGY

This chapter focuses on the research design and methodology. Section 3.1 provides an overview of the fraud classification problem and proposes a fraud detection model. Section 3.2 evaluates various classifiers and discusses why SVM was selected to build the fraud model (see Chapter 5 for more information on the classifiers used in this study). Section 3.3 describes approaches to feature selection, indicating how input features were selected, and leads into Section 3.4, which discusses the research hypotheses and choice of variables. Section 3.5 outlines the procedure that we adopted for testing the hypothesis. Section 3.6 briefly describes the sample and data collection (see Chapter 4 for more information on this). Section 3.7 discusses the preprocessing of such data. Section 3.8 presents the baseline approach, preliminary baseline results and how the classifier was trained (see Chapter 5 for more information on this). Section 3.9 discusses the measures used to test the accuracy of fraud detection model. Finally, Section 3.10 presents some concluding observations.

3.1 FRAUD CLASSIFICATION PROBLEM

Detection of fraud can be considered a classification problem: Is this annual report a fraudulent annual report or a non-fraudulent annual report? Previous studies in accounting and finance that have considered classification problems have focused on modeling auditors’ opinions as qualified or unqualified (Gaganis et al., 2007), detecting
fraud (Persons, 1995; Fanning and Cogger, 1998; Lee et al., 1999; Kaminski et al., 2004), predicting bankruptcy (Tam and Kiang, 1992; Zhang et al., 1999; Anandarajan et al., 2001; McKee and Lensberg, 2002; Pendharkar, 2005), predicting financial distress (Ohlson, 1980), detecting earnings management (Peasnell et al., 2000), assessing the litigation risk associated with audit clients (Berton, 1995), detecting fraudulent insurance claims (Viaene et al., 2002), modeling credit risk (Doumpos and Pasiouras, 2005), modeling stock returns (Hellstrom and Holmstrom, 2000), predicting acquisition targets (Pasiouras et al., 2005) and so forth.

Until recently, most researchers have based their classification models on statistical models such as logistic regression (Persons, 1995; Bell and Carcello, 2000), linear discriminant analysis (Fanning and Cogger, 1998), and probit analysis (Dopuch et al., 1987; Lennox, 2000). More recently, researchers have been using artificial intelligence (AI), data mining, and machine learning techniques to model problems in the domains of accounting and finance. This shift may be attributable to limitations of some of these earlier models and to opportunities for expanding applications of AI technologies in the domain of accounting, particularly auditing and assurance.

Drawing on the field of AI, some of the previous 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. Some of the other fraud detection models used variations of logistic regression (Beasley, 1996; Summers and Sweeney, 1998; Abbot et al., 2000; Spathis, 2002) and probit analysis (Hansen et al., 1996; Beneish, 1999).
This dissertation proposes a methodology to proactively detect fraud by examining qualitative content of annual reports using Natural Language Processing (NLP) tools. The methodology employs machine learning techniques to build an automated classifier that can predict the likelihood of fraud. This study proposes a fraud detection model that was built using Support Vector Machines (SVM), a supervised machine learning technique that incorporates domain knowledge while the learning task is being undertaken from a set of positive and negative examples provided in the training set. This research attempts to explore linguistic features that differentiate fraudulent annual reports from non-fraudulent annual reports. These features, also called predictor (independent) variables, are discussed in detail under the section ‘Hypotheses and Variables.’

Once the learning was successful, SVM was able to successfully classify as fraudulent or non-fraudulent unlabelled annual reports in the testing dataset that did not belong to the training sample. The correctness of fraud predictions was then evaluated against the correct fraud classes of the testing dataset. Several standard evaluation measures such as accuracy, precision, recall, and F-measures were used to test the prediction accuracy of the fraud classification model (Manning and Schutze, 1999). These evaluation measures presuppose that each document (annual report) belongs to only a single class (fraudulent or non-fraudulent). The fraud classifier was also trained on pre-fraud and post-fraud data of fraudulent companies to detect early warning signs of fraud.

The methodology presented in this dissertation differs from earlier fraud detection studies using AI techniques as well as non-AI techniques with respect to input vector selection. Most prior studies have selected quantitative information such as financial ratios and metrics as the input vector. Unlike these earlier studies, this study looked at the
qualitative factors such as tone, readability index, etc. to assess the likelihood of fraud. In addition, in this research we carried out an in-depth examination of the qualitative (non-quantitative) content of annual statements in terms of both content and presentation style, unlike some earlier studies where they looked at only one sub-section of the annual report to predict bankruptcy, companies’ future viability, company performance, and firms’ environmental performance.

3.2 CHOICE OF CLASSIFIER - SVM

Assigning predefined categories to text documents by means of supervised learning has been gaining momentum as a key technology for text mining in different research fields. Many different automatic classification techniques have been proposed by researchers such as Naïve Bayes (NB), Rocchio classifier (Rocchio, 1971), Support Vector Machines (Cortes and Vapnik, 1995), decision trees, neural networks, maximum entropy, k-Nearest Neighborhood (k-NN) and so on for text and document classification applications. In this study, a fraud detection model is built using the SVM classifier. The weaknesses and strengths of various statistical and AI machine learning classification techniques that have been considered by researchers to address similar problems were evaluated before selecting SVM. This section presents a brief overview of the SVM classifier, pointing out why SVM was selected for fraud detection (see Chapter 5 for more information on SVM).

Support Vector Machine is a supervised machine learning technique that is based on statistical learning theory. In machine learning, the goal is to develop classifiers from examples. The SVM algorithm learns by example to classify objects into a fixed number
of predefined categories. In this study, SVM was trained to recognize fraudulent annual reports by examining hundreds of fraudulent and non-fraudulent annual reports. This approach of using examples when the examples are input-output pairs is called supervised learning (Cristianini and Shawe-Taylor, 2000). The examples of input-output pairings used in supervised learning are referred to as training data.

SVMs are based on the Structured Risk Minimization (SRM) principle drawn from computational learning theory (Vapnik and Cheronekis, 1974). SRM is an inductive principle for model selection that provides a trade-off between hypothesis space complexity and the quality of fitting the training data, which guarantees the lowest true error on an unseen and randomly selected test example. SVMs determine a hyperplane in the feature space that best separates positive from negative examples.

Classifiers based on machine learning techniques suffer from an endemic problem of overfitting especially in the case of small datasets. Overfitting is the phenomenon by which a classifier is tuned to the “contingent” characteristics of training data, as opposed to the “constitutive” characteristics of the class and typically produces good results on the training dataset but not on the test dataset. Increasing the size of the training set or using the technique of cross-validation can resolve the problem of overfitting to some extent. Some classifiers are more sensitive to the problem of overfitting than others. For example, it is widely known that the phenomenon of overfitting is more common in neural networks and Bayesian approaches (Witten and Frank, 2005). Overfitting is not as common with SVM as with the other classifiers because SVMs are based on structural risk minimization; whereas classifiers such as traditional neural networks are based on empirical risk minimization.
Across a wide spectrum of applications, SVM boasts remarkable empirical results, outperforming other classifiers, particularly in problems that deal with large feature sets or high-dimensional datasets (Bradley et al., 1998; Yu et al., 2003). Moreover, SVMs have a remarkable ability to learn, independent of the dimensionality of the feature space. SVM is the most successful machine learning method in NLP, and has been shown to yield the highest levels of accuracy in comparison to other learning algorithms in classification and regression applications (Joachims, 1997, 1998; Dumais et al., 1998). Besides text classification, SVM has been applied to other NL tasks such as chunking (Kudo and Matsumoto, 2001), dependency analysis (Kudo and Matsumoto, 2002) and so on.

SVMs have also been shown to excel in binary classification problems more than in multi-class text classification problems. In contrast to other classifiers, SVMs produce improved generalization accuracy, even in the presence of many features (Joachims, 2000). Brown et al. (1999) successfully applied SVMs to the problem of classifying unseen genes. Another advantage of SVMs is that they consistently achieve good performance in problems with dense concepts and sparse instances, avoiding catastrophic failures observed with other classifiers. All this makes SVMs better suited than other classifiers for learning that is specific to the domain problem of fraud classification.

3.3 FEATURE SELECTION

Feature selection is a common technique that is used in machine learning to select a subset of relevant features from available potential candidate features. This involves finding independent, indispensable, and strongly relevant features that complement each
other and contribute to defining the unknown concepts. The irrelevant features that do not contribute to defining the unknown categories are eliminated. The weakly relevant features can also be removed if they contain redundant information that has already been included with other features. A relevant feature is defined as one whose removal adversely affects the accuracy or performance of classifier.

Feature selection for text classification is a well-researched problem in fields ranging from life sciences to linguistics. In the context of supervised learning, feature selection techniques can be categorized into two approaches: filter approach and wrapper approach (Kohavi and John, 1997). A filter approach selects variables by ranking them with correlation coefficients, independent of the choice of the predictor. In contrast, wrapper approach assesses subsets of variables according to their usefulness to a given predictor. Some of the core methods used for feature selection are document frequency, information gain, mutual information, and chi-square. Document frequency counts the number of documents containing the feature. Information gain is the number of bits of information obtained for category prediction given a feature. Mutual information measures mutual dependence of the two variables. Chi-square measures the lack of independence between a term and the category.

Feature selection methods can also be broadly categorized as either redundancy reducing methods or aggressive feature space reduction methods. Feature space reduction methods have repeatedly resulted in performance gain with little accuracy loss but results are not encouraging at high performance levels (Soucy and Mineau, 2001). Clustering has also been used for feature selection where a cluster centroid of a group of similar variables is considered a feature. However, clustering is usually associated with unsupervised
learning. Bartell et al. (1992) have used latent semantic indexing with singular value
decomposition to reduce the dimensionality of the feature space. Moore et al. (1997) have
used association rules and principal component clustering for feature selection. Rozsypal
and Kubat (2001) have used machine learning techniques such as genetic algorithms to
identify features relevant to their domain of study. Other researchers have used hybrid
approaches such as genetic algorithms and mutual information to address dependency
and redundancy among features (Koller and Sahami, 1996; Das, 2001; Xing et al., 2001;
Torkkola, 2003). Techniques such as grid search have been also used with some success
(Lin et al., 2006).

Feature selection can also be implemented using two types of search strategies, which
have been found computationally advantageous and robust against overfitting: forward
selection approach and backward elimination approach (Guyon and Elisseeff, 2003). In
forward selection, one starts with an empty set and adds features one at a time. In
contrast, in backward elimination, one starts with a feature set containing all features and
eliminates one at a time (the least promising one). Other commonly used search strategies
are forward stepwise selection, backward stepwise elimination, and random mutation. In
forward stepwise selection, one starts with an empty set and adds or removes features one
at a time. In backward stepwise elimination, one starts with a feature set containing all
features and adds or removes features one at a time. In random mutation, one starts with a
feature set containing randomly selected features, and adds or removes randomly selected
features one at a time and stops after a given number of iterations.

In order to create effective and high performance model for fraud detection, it is
important to distinguish features that improve classifier performance from features that
make no difference in the classifier performance or adversely affect its performance. If all
the features are considered at once, it is difficult to identify features relevant to fraud
detection or distinguish between strong and weak predictors of fraud. Moreover,
inclusion of all possible features can deteriorate the performance of the classifier.
Therefore, a forward stepwise selection approach to feature selection was chosen where
features were added one by one, inspired by informed reasoning gathered from the
literature review and intuition. The objective was to identify features more efficiently and
thus to uncover clues hidden in the qualitative text of annual reports that can predict
fraud.

To construct a feature set that captured the most significant aspects of the data, more
linguistically-oriented variables representing fraud cues were explored using the forward
stepwise selection approach. The selection of final feature set was based on the
discrimination power of the features.

3.4 HYPOTHESES AND VARIABLES

This research examines both the verbal content and presentation style of the qualitative
content of companies’ annual reports to explore linguistic features that distinguish
fraudulent annual reports from non-fraudulent annual reports. Our initial baseline results
show that a classifier based on a full ‘bag of words’ approach was able to achieve only
52.17 percent fraud classification accuracy. We build a fraud detection model that
achieves higher levels of accuracy than the baseline full ‘bag of words’ approach.
Specifically, this research attempts to find out if indicators of fraud such as keywords,
tone, sentiment, readability index, complexity, type of voice, uncertainty markers, and
other linguistic information derived from the qualitative portions of the corporate annual reports can be used to proactively detect fraud.

We now present the hypotheses and then present the predictor (independent) variables that were used to construct the feature set. Initially, the feature set consisted of the features that best represented the documents. As the classifier converged into higher levels of accuracy, this feature set evolved during training and the final set consisted of a subset of this and other novel features. The predictor variables are categorized into two categories: content features and presentation features. Content features relate to verbal content of the qualitative portion of the annual report such as keywords and TFIDF-weighted tokens, whereas presentation features such as tone, sentiment, readability index, sentence complexity, type of voice and uncertainty markers relate to its presentation style. This section concludes with a discussion of dependent and control variables.

3.4.1 HYPOTHESES

In cases of fraud, companies try to misrepresent information and therefore employ writing techniques that differ from the companies that do not commit fraud.

Our main hypothesis, stated above, is of interest as it is believed that, in order to conceal fraud, companies resort to many creative techniques to proactively construct a meaning rather than reveal “what was there.” Companies use both selective accounting language and presentation style features to hide fraud and therefore text contains more diverse and dense information than numbers do. In order to test this main hypothesis, we develop five related sub hypotheses that are discussed next.
H1: *The greater the use of complex sentential structures in the qualitative content of a company’s annual report, the greater the likelihood that there will be fraud.*

Hypothesis H1 is of interest as it is believed that companies consciously employ means to obscure real information that might reveal fraud by presenting it in a convoluted manner. Conversely, clear succinct language is employed to convey information more clearly and this requires short and direct sentential constructs. For H1, the null hypothesis that the count of complex sentential structures indicated by features such as ambiguity index and type-token ratio is similar in both fraudulent and non-fraudulent annual reports was tested.

H1a: *The more difficult it is to read and understand a company’s annual report, the greater the likelihood that there is fraud.*

Related to hypothesis H1, we define another sub hypothesis, H1a. This hypothesis is of interest as it is believed that, in cases of fraud, companies deliberately make annual reports difficult to comprehend. In order to disguise evidence of fraud, management intentionally uses longer sentences and more difficult words so that it is harder to grasp and understand them. There is wide consensus among researchers that the annual report is nothing but a highly sophisticated product of the corporate design environment. Conversely, to make reports easier to understand, management makes use of shorter sentences with simple words to improve the readability of these reports. For H1a, the null hypothesis, that readability scores are similar in both fraudulent and non-fraudulent annual reports was tested.
H2: The greater the use of negative words in a company’s annual report, the greater the likelihood that there is fraud.

Hypothesis H2 is of interest as it is believed that management consciously chooses selective accounting language to conceal fraud. The annual report typically reflects on a company’s strategy and, when things are not going well, its numbers do not change but qualitative content changes. For H2, the null hypothesis, that negative and positive words are used similarly in the qualitative content of fraudulent and non-fraudulent annual reports, was tested, i.e., the null hypothesis stated that there exists no relationship between the polarity of tone and the outcome of fraud.

H3: The greater the use of passive voice in a company’s annual report, the greater the likelihood that there is fraud.

Hypothesis H3 is of interest as it is believed that, in cases of fraud, management consciously tries to shift responsibility away from itself and makes more use of passive voice sentences than active voice sentences. Conversely, in order to take credit for positive outcomes and attribute these outcomes to its own actions, management will use active voice sentences. For H3, the null hypothesis, that there is no relationship between type of voice and outcome of fraud was tested.

H4: The greater the use of uncertainty markers in a company’s annual report, the greater the likelihood that there is fraud.

Hypothesis H4 is of interest as it is believed that, in cases of fraud, management deliberately employs more uncertainty markers to make reports ambiguous. Management
may rationalize this act by arguing that it is better to introduce uncertainty than make false promises. Conversely, to provide a definite and clear picture, management will avoid using uncertainty markers. For H4, the null hypothesis, that use of uncertainty markers is similar in the qualitative content of fraudulent and non-fraudulent annual reports was tested.

### 3.4.2 PREDICTOR VARIABLES

**Term Frequency (TF):** Term Frequency is the raw frequency count of the terms within the document. Before calculating term frequency, a stop list of commonly used English words that do not carry semantic meaning was applied. All terms whose count was less than or equal to three were eliminated to boost the classifier performance without sacrificing its accuracy. This data preprocessing is explained in detail under the section ‘Preprocessing dataset.’

**Weighted-Term Frequency Inverse Document Frequency (Weighted-TFIDF):** Inverse Document Frequency (IDF) is the negative log of the number of documents within which the term appears, divided by the total number of documents. A term-document matrix is created that represents every document and content word in a corpus. Content words that appear in every document are discarded, since such words do not contribute to discrimination between documents. Similarly, content words that appear in only one document are also discarded, as such words do not tell about relationships across documents.

In a term-document incidence matrix, ones indicate the presence of a term in the document and zeros represent the absence of a term in the document. For each non-zero
term/document, term weighting, including local weighting, global term weighting, and normalization, is applied. According to local weighting, content words that appear several times in a document are probably more meaningful than content words that appear just once and thus are given a greater local weight. Global term weighting applies to the set of all documents in the collection and reflects the fact that words that appear in a small handful of documents are likely to be more interesting and of greater discriminating value than the words that are widely distributed across the document collection. The final step to weighting is called normalization, where every document is scaled so that it has equal significance and large documents with many keywords will not bias the results. There is extensive literature demonstrating that frequencies of words convey information regarding their importance and content captures otherwise hard-to-quantify concepts. (see, for example, Zipf 1929, 1949; Luhn 1957; Iker 1974; Weber 1990; Gangolly and Wu 2000; Hand et al., 2001; Uzuner and Katz, 2005a).

**Voice:** Voice is measured by counting frequencies of active voice and passive voice sentences. It is believed that, in the presence of fraud, the management consciously tries to shift responsibility away from itself and makes greater use of passive voice to introduce uncertainty (or at least ambiguity) as to responsibility. This is evident by use of truncated passives where the agent is left out altogether. Conversely, in order to take credit for positive outcomes and attribute these outcomes to its own actions, the managements use active voice in the annual reports.

**Frequency of uncertainty markers:** Uncertainty markers (also known as hedge words or modal verbs) include words such as “shall,” “may,” “probably,” “possibly,” “might,” etc. Uncertainty markers have been used in studies to detect style and expression of
literary narratives (Glover and Hirst, 1996; Uzuner and Katz, 2005b). It is believed that in cases of fraud, the management deliberately employs more uncertainty markers to render the reports ambiguous. The management can rationalize this act by the argument that it is better to introduce uncertainty than to make untrue disclosures.

**Readability Index:** A readability index (RI) takes into account surface features of the text and is designed to measure the understandability of a text. Linguistic measures such as sentence length and word length have been widely used to indicate semantic and syntactic levels of difficulty in literary texts. In this research, a RI is used to compare readability of annual reports for fraudulent and non-fraudulent companies. Prior research on measuring the readability of corporate annual reports shows mixed results.

Smith and Taffler (1992) compared annual reports of failed firms with non-failed firms using Lix and Flesch readability tests and found that corporate failure is significantly correlated with more difficult syntax. Subramanian et al. (1993) compared annual reports of good performing corporations with poor performing corporations and concluded that annual reports of good performers are significantly easier to read than those of poor performers. However, Baker and Kare (1992) found that there is no relationship between the Flesch score and the profitability of a company. Similarly, Courtis (1986) found no systematic relationship between the readability score of a company’s annual report and a company’s level of risk and return. The difference in the results of the relationship between readability and corporate characteristics may be due to the fact that some studies sample the whole text of annual reports (Smith and Taffler, 1992) whereas other studies test only selected sections of the annual reports.
There are many tools to compute a readability index such as ‘Gunning fog index,’ ‘Flesch-Kincaid Readability Test,’ ‘SMOG Index,’ and ‘Coleman-Liau Index.’ The Flesch Reading Ease Score (Flesch, 1949) is a frequently used and highly recognized measure to analyze readability of a text. Higher scores on the Flesch Index indicate that text is easier to read whereas lower scores indicate a greater reading difficulty. As a convention, Flesch scores from 0-30, 30-50, and 50-60 are interpreted as very difficult, difficult, and fairly easy to read, respectively. Flesch uses two core measures; sentence length and word length to diagnose the writing style and, in particular, only one aspect of style, that is, difficulty. The length of a sentence is measured in the number of words that consist of three or more syllables after stop words have been removed from the sentence. The Flesch Reading Ease Score is calculated as:

\[
\text{Flesch Index} = 206.835 - 1.015 \times \text{ASL} - 84.6 \times \text{ASW}
\]

where ASL = average sentence length (the total number of words divided by the total number of sentences), and

\[
\text{ASW} = \text{average number of syllables per word (the total number of syllables divided by the total number of words)}.
\]

**Tone:** The tone of a text refers to the emotional attitude it expresses. A text’s tone can be classified as happy, sad, optimistic, pessimistic, positive, negative, objective, subjective and so on. The detection of attitude, sentiment, polarity and subjectivity in text has received increased attention over the past several years. Henry (2005) found that there is a relationship between tone and investor reaction to earnings announcement. Abrahamson and Amir (1996) examined the relationship between negative words in the chairman’s
letter and market returns. Smith and Taffler (2000) described the relationship between negative words in the CEO letter and firm failure.

In this study, tone is defined as positive (optimistic) or negative (pessimistic). Tone is measured in terms of its polarity, i.e., as the frequency count of positive (optimistic) and negative (pessimistic) words, scaled by the total number of words in an annual report. The words that are defined as negative (pessimistic) and positive (optimistic) are listed in Chapter 6. This list was created from the literature review (Abrahamson and Amir, 1996; Smith and Taffler, 2000; Hand et al., 2001; Henry, 2005) augmented by additional words specific to fraud. The negative and positive categories are mutually exclusive so that no selected word or phrase falls in both categories. Each annual report was represented as a vector of scaled frequency counts of positive and negative words.

3.4.3 DEPENDENT VARIABLE

**Fraud:** In this study, the dependent variable is fraud. Fraud is a dichotomous variable; it was measured by the presence (fraudulent or 1) or absence (non-fraudulent or 0) of fraud.

3.4.4 CONTROL VARIABLE

**Size:** The impact of company size that results from larger companies attaching more importance to the annual report as an external communication device and their annual reports being characterized as having greater public visibility is well established in the literature (Fombrun and Shanley, 1990). In order to account for this overrepresentation of larger companies, company size was used as a control variable.
3.5 PROCEDURE OF HYPOTHESIS TESTING

In this study, we considered five hypotheses that we evaluated using a chi-square test. The chi-square is a standard test of statistical significance for hypothesis testing. The results of the chi-square tell us if observed results are significantly different than would expected be due to chance. A research hypothesis is defined based on the expectation that there is a relationship between two variables, X and Y. Then a null hypothesis is defined by negating the research hypothesis, i.e., that there is no relationship between X and Y. The results of chi-square test indicate if the null hypothesis can be rejected. The null hypothesis is rejected if the p-value associated with the computed chi-square statistic is less than the ‘alpha’ value at a chosen level of confidence (\(\alpha = .05\)) and it is concluded that observed results are statistically significant. In other words, 95 out of 100 times the relationship mentioned in research hypothesis is real and not due to chance fluctuation.

Next, we briefly describe the procedure for testing the hypothesis and calculating the value of chi-square. The computation of the chi-square statistic starts with the construction of a contingency table. The cells of this contingency table represent the observed frequency of the data in all possible categories. Tables 3.1 and 3.2 present two contingency (2X2) tables. The first table provides a generic template that we used for testing our hypotheses whereas the second table provides a template for an actual hypothesis relating to voice.
TABLE 3.1
Template of a 2X2 Contingency Table for Testing a Fraud Hypothesis

<table>
<thead>
<tr>
<th>Features</th>
<th>Feature A</th>
<th>Feature B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fraudulent Annual</td>
<td>Number of times feature A is observed in fraudulent</td>
<td>Number of times feature B is observed in fraudulent</td>
</tr>
<tr>
<td>Reports</td>
<td>annual reports</td>
<td>annual reports</td>
</tr>
<tr>
<td>Non-Fraudulent</td>
<td>Number of times feature A is observed in non-</td>
<td>Number of times feature B is observed in non-</td>
</tr>
<tr>
<td>Annual Reports</td>
<td>fraudulent annual reports</td>
<td>fraudulent annual reports</td>
</tr>
</tbody>
</table>

TABLE 3.2
Template of a 2X2 Contingency Table for Testing a Fraud Hypothesis Relating to Voice

<table>
<thead>
<tr>
<th>Features</th>
<th>Feature (Active Voice)</th>
<th>Feature (Passive Voice)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fraudulent Annual</td>
<td>Number of times active voice sentences are observed</td>
<td>Number of times passive voice sentences are observed</td>
</tr>
<tr>
<td>Reports</td>
<td>in fraudulent annual reports</td>
<td>in fraudulent annual reports</td>
</tr>
<tr>
<td>Non-Fraudulent</td>
<td>Number of times active voice sentences are observed</td>
<td>Number of times passive voice sentences are observed</td>
</tr>
<tr>
<td>Annual Reports</td>
<td>in non-fraudulent annual reports</td>
<td>in non-fraudulent annual reports</td>
</tr>
</tbody>
</table>

The next step is to compute the expected frequency for each cell of this table. The formula for calculating expected cell frequency is as follows:

\[
\text{Expected Cell Frequency of the Desired Cell} = \frac{(\text{Row Total} \times \text{Column Total})}{N}
\]

For example, in Table 3.1, the expected frequency for the number of times feature A is present in fraudulent annual reports is:
\[ = \frac{\text{(Total number of times feature A occurs in both fraudulent and non-fraudulent annual reports} \times \text{Total number of times features A and B occur in fraudulent annual reports)}}{N} \]

Once the expected cell frequency for all the cells in the table is calculated, the next step is to compute the difference between the observed cell frequency and the expected cell frequency for all the cells in the table. In the last step, we square the difference between the observed and expected frequency and divide it by the expected cell frequency for each cell. The value of chi-square statistic is computed by adding all the values obtained in the last step. In other words, chi-square is the sum of the squared difference between the observed and the expected frequency divided by the expected frequency for all the possible categories.

The formula for computing chi-square is given by:

\[ \chi^2 = \sum_{\text{cells}} \frac{(O - E)^2}{E} \]

where \( O \) = observed frequency

\( E \) = expected frequency

After computing the chi-square statistic, we assess its significance level and decide whether the null hypothesis can be rejected or not. In order to determine the significance level of the chi-square statistic, we first determine the ‘degrees of freedom.’ In the case of a contingency table, degrees of freedom are computed by multiplying the number of rows minus one by the number of columns minus one.
The formula for calculating degrees of freedom is given by:

\[ \text{Degrees of Freedom} = (\text{Number of Rows} - 1) \times (\text{Number of Columns} - 1) \]

The degrees of freedom also determine the shape of the chi-square distribution. Once we know the p-value of the chi-square statistic, we compare it to the ‘alpha’ value at the chosen level of confidence \( \alpha = .05 \) and we reject the null hypothesis if the p-value is less than the alpha value. It should be noted that the chi-square statistic is known not to be reliable if the value of any cell in the contingency table is less than five. Furthermore, chi-square makes two fundamental assumptions: (1) all the observations in the contingency table are independent of each other, and (2) the sample size is large enough such that the expected frequency in each cell of the contingency table is greater than or equal to 5.

During hypothesis testing, we need to be aware of the risk of type I and type II errors. Type I error occurs if there is no relationship between two variables X and Y and it is concluded that there is a relationship. On the other hand, type II error occurs if there is a relationship between two variables, X and Y, and it is concluded that there is no relationship.

### 3.6 SAMPLE AND DATA COLLECTION

A sample of 126 fraud (405 fraud firm years) and 622 no-fraud (622 no-fraud firm years) companies was selected from 1993 to 2006 in three stages (see Chapter 4 for detailed information on sample and data collection). Initially, a comprehensive list of all those US publicly listed companies where fraud had occurred over a 14-year period from 1993 to 2006 was created. LexisNexis and Compustat, via Research Insight, were the primary
sources of gathering all fraud company data. Many empirical studies on financial statement fraud have used the issuance of Accounting and Auditing Enforcement Releases (AAERs) issued by the SEC as a proxy for financial statement fraud. Even though AAERs provide an objective way of identifying publicly traded companies that have been accused of financial statement fraud, many companies accused in AAERs reach a settlement with the SEC, without admitting or denying the allegations. In this study, a list of companies that have been accused in AAERs but not included in the above sample was collected to create an additional testing dataset.

The time period from 1993 to 2006 was selected for two reasons. First, 10-Ks are electronically available in the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) database starting in 1994. Second, a majority of high profile accounting scandals occurred during this period, which eventually led to the enactment of the Sarbanes-Oxley legislation in 2002. The time period ends at 2006 because the annual financial reports beyond fiscal year 2006 were not yet available at the time of the study.

A total of 140 companies where fraud had occurred and been discovered were identified during this time period. Out of these 140 companies, a total of 126 companies that had filed their 10-Ks with the SEC and whose 10-Ks are electronically available for download from the EDGAR online database were selected.

A total of 622 US publicly listed corporations where fraud had never been reported from 1993 to 2006 and whose 10-Ks are available for electronic download from the EDGAR database were also selected. In this study, a peer set for each fraud company as opposed to a matched-pair dataset was selected. This choice of sample approximates a more realistic scenario of infrequency of fraud. The inclusion of multiple peer no-fraud firms
for each fraud firm also protects against the bias of overstating the importance of explanatory variables that often occurs in a matched-pair approach (Zmijewski, 1984; Lee et al., 1999).

Furthermore, selection of no-fraud companies was done on the basis of two rules. First, for each fraud company, multiple no-fraud companies were selected for the fraud year. This was done to ensure that changes in GAAP or any other accounting rules and regulations did not account for the differences in the findings. Second, for each fraud company, multiple no-fraud companies were selected from the same industry and in the same size range as defined by 2-digit SIC codes. For example, if one fraud company falls under communications (48), then approximately five no-fraud companies were selected from the communications industry that are within the same percentage (10-20%) of total assets or sales of the fraud company (see Chapter 4 for detailed information on no-fraud sample).

Moreover, for both sets of fraud and no-fraud companies, the original 10-Ks filed with the SEC were collected and not the restated 10-Ks. The original 10-Ks were selected because a restatement of a financial statement is typically created to correct the previous financial statement for intentional/unintentional errors and accounting irregularities. Restatements represent an acknowledgement by the firm that prior financial statements were not in accordance with GAAP (Palmrose and Scholz, 2004). In order to identify symptoms of fraud and proactively detect fraud, we needed to examine and analyze original 10-Ks and not the restated 10-Ks.
In addition, 8-K filings that are available for selected companies during this time period were also collected. Form 8-K is the “current report” companies must file with the SEC to let shareholders know about major events such as auditor changes, restatements, significant acquisitions, bankruptcy, large write-offs, major restructuring changes and so on. These episodic 8-K filings are collected as they sometimes contain information disclosures that are not otherwise reported and may provide additional insight.

### 3.7 Preprocessing of the Dataset

The preprocessing stage basically addresses the question as to how text should be represented. Here, raw data (company filings) are transformed into a compact representation, after a number of initial operations are performed. Depending on the type of variables, these steps were adjusted.

Function words (stop words) such as articles (a, an, the), conjunctions (and, or), and prepositions (in, of, at) were removed. The stop word list consists of commonly used English words that do not carry semantic meaning and thus do not have a significant impact on the classification of the document. Applying the stop list reduces noise in the data and improves performance. However, auxiliary verbs (to be, to have, can, may, must, shall, will, want, etc.) and conjunctions (not) were not removed since these words were required to analyze other linguistic features central to our hypotheses. The stop list was adjusted before it was applied to the data. Once stop words were removed, the resulting dataset was an abbreviated version of the data containing only content words and some stop words. The resulting dataset allowed for the generation of the initial term-
document matrix, where each document was transformed into a frequency vector of all
words that appear in the document.

Stemming was not performed as inflected variants of base forms have very different
meanings in the domain of accounting. For example, ‘security’ means the instrument
representing ownership share or bond instrument where as ‘secure’ means safe or
protected. Similarly, ‘tender’ means an offer or bid made for acceptance but ‘tend’ means
to care for; ‘arbitrage’ means simultaneous purchase and sale of the same securities in
different markets but ‘arbitrary’ means subject to individual judgment. This is consistent
with established practice in existing research (e.g., Chen et al., 1995; Garnsey, 2006).

3.8 TRAINING THE CLASSIFIER

In this study, the classifier was trained using 10-fold cross-validation. For the first
version of the fraud detection dataset, the corpus was divided into 925 training documents
and 102 test documents (see Chapter 5 for more information on training and testing
datasets). First, the classifier was trained with a full ‘bag of words’ approach to get
baseline results. Then, one processing feature was changed at a time in order to compare
the contribution of the individual processing features in its performance.

**Baseline:** Using a full ‘bag of words’ approach, all document tokens were submitted into
the Bow (also known as Rainbow) classifier system, which yielded 52.17% classification
accuracy. Rainbow is a statistical text classification program developed by Andrew
McCallum (1996). It is based on the Bow library and uses the Naïve Bayes (NB)
algorithm for text classification. The basic idea in the NB approach is to use the joint
probabilities of words and categories to estimate the probabilities of categories when a
document is given (McCallum and Nigam, 1998). Using the same approach, baseline results were obtained with the SVM classifier, the main classifier used in the study (see Chapter 5 for more information on baseline results obtained with these classifiers).

**Learning:** The learning algorithm used for fraud detection is a SVM classifier. A 10-fold cross-validation was conducted and its results were averaged. Specifically, the training set was partitioned into ten sets as discussed earlier and, circularly, nine sets were used for training and the leftover set was used for validation. Once an optimal threshold was determined through cross-validation, a final SVM model was trained using the entire training set for that class.

**Chi-Square:** In this study, a chi-square method was used to select features that show statistically significantly differences between the positive and negative documents. Chi-square feature selection has been shown to not only reduce the feature space effectively by reducing the noise introduced in the classifier, but also to improve performance of the classifier at the same time. A 2x2 contingency (cross-classification) table was constructed for each feature and the p-value was computed using chi-square. The p-value was used to decide whether or not the null hypothesis could be rejected. Features with p values less than the ‘alpha’ value for the chosen level of confidence ($\alpha = .05$) were used as input to the classification algorithm. These data also indicated if the likelihood of detecting fraud is related or not related to this feature. For instance, if observed results (p-value is less than alpha value) indicate that the increase in classifier prediction accuracy is statistically significant with a variable at chosen levels of confidence ($\alpha = .05$ or $\alpha = .01$), the null hypothesis (no relationship between this type of variable and prediction accuracy) would
be rejected. Using probability \( p < .05 \) indicates that 95 out of 100 times, this relationship is real and not random.

3.9 TESTING THE CLASSIFIER

The F-measure, defined as the weighted harmonic mean of precision and recall, was used to measure the performance of the fraud detection model in this study. The F-measure typically gives equal weight to both precision and recall and computes the point where precision and recall are equal. The formula for computing the F-measure is:

\[
F\text{-}measure = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

Precision for class X is defined as the number of correct predictions of X divided by the total number of predictions of X. Recall for class X is defined as the number of correct predictions of X divided by the total number of actual instances of X in the dataset. The formulae for computing precision and recall are:

\[
\text{Precision} = \frac{(\text{Total number of true positives})}{(\text{Total number of true positives}) + (\text{Total number of false positives})} \times 100
\]

\[
\text{Recall} = \frac{(\text{Total number of true positives})}{(\text{Total number of true positives}) + (\text{Total number of false negatives})} \times 100
\]

where true positives = fraudulent report correctly classified as fraudulent

false positives = non-fraudulent report incorrectly classified as fraudulent

false negatives = fraudulent report incorrectly classified as non-fraudulent
true negatives = non-fraudulent report correctly classified as non-fraudulent

There are two methods to compute the F-measure: micro-averaging and macro-averaging. In micro-averaging, precision and recall are computed first for all the categories and then the values of precision and recall are used to calculate the F-measure. In contrast, macro-averaging involves first computing the F-measure for each category and then averaging them. Typically, F-scores using micro averaging are higher than their corresponding macro-average F-scores and recall is favored over precision in order to avoid losing any information that is useful as higher precision scores can be achieved by further filtering. In this study, we use the formulae described in this section to calculate recall, precision and the F-measure.

3.10 CONCLUSION

This chapter presented a methodology that combines aspects of fraud detection research in accounting with computational linguistic tools to aid in fraud detection. This research takes a natural language processing approach to examine both the content and presentation style of qualitative textual content of publicly available financial statement data by extracting linguistic features from these statements. These linguistic features were then passed in an input vector to a SVM classifier and, using 10-fold cross-validation, a classifier was trained on annual reports of fraudulent and non-fraudulent companies for the period 1993 to 2006. The output vector of the classifier consisted of the results of the classification, i.e., to which class did an annual report belong: Is it a fraudulent 10-K or a non-fraudulent 10-K? The trained classifier was then used to classify new data.
CHAPTER 4
SAMPLE AND DATA COLLECTION

This chapter provides data definitions, the sample selection process, and the data collection methodology for both fraudulent and non-fraudulent companies. In Section 4.1, data definitions and the rationale behind them are articulated. Section 4.2 discusses the sample construction, the data collection methodology, and describes the sample that was finally selected for building the fraud detection model. Section 4.3 summarizes the sample design.

4.1 DATA DEFINITIONS

The dataset for this study consists of the annual reports (10-Ks) of both fraud and no-fraud companies identified during the sample period (1993-2006). The sample period includes the fraud period, the pre-fraud period and the post-fraud period. Since the fiscal year varies from company to company, a company’s reporting period was taken into consideration when defining the fraud period and collecting its 10-Ks.

Fraud Dataset

The fraud dataset consists of companies that were accused of fraudulent financial reporting in the period from 1993 to 2006, i.e., fraud that had affected 10-Ks (annual reports) through material manipulation, misrepresentation or failure to disclose material
facts. Specifically, a company was included if it was alleged of violating Rule 10(b)-5 of the Securities Exchange Act of 1934 and, subsequently, sufficient evidence of fraud was found to corroborate the original allegations. Rule 10(b)-5 requires the intent to deceive, manipulate, or defraud. Thus, cases where a company was alleged to accept kickbacks, violate Foreign Corrupt Practices Act, participate in a price fixing scheme, violate antitrust laws, conduct wire fraud, issue a fraudulent prospectus, or commit fraud on registration statements were excluded from the fraud dataset.

Fraudulent companies were identified using Lexis-Nexis, Compustat via Research Insight, the Wall Street Journal (WSJ) Index and Accounting and Auditing Enforcement Releases (AAERs) issued by the Securities Exchange Commission (SEC) for the period 1993 to 2006. Many empirical studies on financial statement fraud have used the issuance of AAERs as a proxy for financial statement fraud. Even though AAERs provide an objective way of identifying publicly traded companies that have been accused of financial statement fraud, many companies accused in AAERs reach a settlement with the SEC, without admitting or denying the allegations. In order to make sure that the fraud dataset did not include non-fraudulent companies, only those AAERs where companies failed to comply with SEC rules that pertain to fraud and had documented evidence of fraud were considered.

No-Fraud Dataset

In this study, a peer set for each fraud company as opposed to a matched-pair dataset was selected. This choice of sample approximates a more realistic scenario of the infrequency of fraud. The inclusion of multiple peer no-fraud companies for each fraud company also
protects against the bias of overstating the importance of explanatory variables that often occurs in a matched-pair approach (Zmijewski, 1984; Lee et al., 1999).

For each fraud company, an attempt was made to select five no-fraud companies. The selection of no-fraud companies was done on the basis of a three-fold criterion of year, industry and size. First, for each fraud company, multiple no-fraud companies were selected for the fraud period. This was done to ensure that the changes in GAAP or any other accounting rules and regulations do not account for the differences in the findings. Second, for each fraud company, multiple no-fraud companies were selected from the same industry as defined by two-digit Standard Industrial Classification (SIC)\(^2\) codes, where available, and North American Industry Classification System (NAICS)\(^3\) otherwise. The third criterion for selection of no-fraud companies was that they had to be within the same size range, i.e. within 10-20 percent of the total assets or sales of the fraud company. For example, if six fraud companies were identified under communications (two-digit SIC code 48), then the same proportion of no-fraud companies (5 percent) that met the size requirement were also selected from this industry for the fraud period.

In addition, in order to avoid recognizing the company’s style, rather than the presence of fraud, we did a comparative study on two no-fraud datasets to evaluate the effects of company’s style on the performance of fraud detection model. For this, we created two

\(^2\) SIC is a coding system that was developed by the U.S. government for classifying industries. SIC system is divided into five levels. Letters A-K indicate main divisions. 2-digit SIC codes indicate major groups within the main divisions. 4-digit SIC codes indicate specific industries within the major groups. 6-digit SIC codes indicate sub-industries within the specific industries. 8-digit SIC codes indicate lines of business within the sub-industries.

\(^3\) NAICS is a new system of coding that was introduced as a replacement for SIC system. However, certain government departments and agencies, such as the SEC still use the SIC codes.
versions of a no-fraud dataset and each was paired with a fraud dataset. As described earlier, the first version of the no-fraud dataset consisted of 10-Ks of no-fraud companies that met the selection criteria. In addition to the 10-Ks of no-fraud companies, the second version of the no-fraud dataset also included 10-Ks for non-fraudulent years of selected fraud companies that were outside the pre-fraud, fraud, and post-fraud periods. Thus, for the second version of the no-fraud dataset, 348 non-fraudulent years of the selected fraud companies were identified that were outside pre-fraud, fraud, and post-fraud periods. We term this set (the second version of the no-fraud dataset) of the 10-Ks as the ‘non-fraudulent reports of mixed companies’ whereas we term the 10-Ks of the first version of the no-fraud dataset simply as ‘non-fraudulent reports.’ It should be noted that in the second set, the same companies were included in both the fraud and no-fraud datasets to control for the company’s style by including non-fraudulent years of fraud companies, whereas in the first set, the two corpora (fraud and no-fraud) included 10-Ks of different companies. For all the experiments that we conducted with the first version of no-fraud dataset, we repeated each of those experiments with the second version of the no-fraud dataset and compared their results. The experiments and discussion of the experimental findings are described in the results section of the dissertation.

Fraud Period

In this study, the fraud period is defined as the time period of alleged fraud for a fraudulent company and one year immediately preceding this alleged fraud period. This is done to ensure that the fraud dataset includes all the 10-Ks for the years when the fraud period.

---

*In the rest of the dissertation we use this terminology to refer to the 10-Ks of the second version of the no-fraud dataset and the ‘non-fraudulent reports’ to refer to the 10-Ks of the first version of the no-fraud dataset.*
company was accused of committing fraud and of issuing materially false and misleading statements. Thus, each year of the fraud period matches the year of the company’s performance when it was alleged to have perpetrated fraud. LexisNexis, Compustat, the WSJ Index, and SEC’s AAERs for the period 1993 to 2006 were the primary sources of gathering all fraud company data. Since, all U.S. publicly listed companies are required to file their 10-Ks with the SEC within 90 days of the end of their fiscal year, the first year of the alleged fraud period typically includes statements for the prior year. In order to avoid confounding of data, the fraud period in this study corresponds to all the years of the company’s performance when it was alleged to be fraudulent. For example, the time period of alleged fraud for Computer Associates is found to range from 1999 to 2002 and during this time it issued (filed) materially false and misleading statements. In this case, the fraud dataset not only includes 10-Ks for its fiscal years ending December 1999, 2000, 2001, and 2002, but also for the fiscal year ending December 31, 1998, and therefore, its fraud period ranges from 1998 to 2002.

Pre-Fraud Period

For fraud companies, 10-Ks are collected for four years prior to the fraud period, which constitutes its pre-fraud period. The inclusion of four years prior to the fraud period was done to be consistent with the literature findings, which indicate that intensity of fraud grows over time and it usually takes an average of 3.02 years before a fraud is exposed (Summers and Sweeney, 1998). Data on the pre-fraud period were collected to detect early warning signs of fraud. In order to detect these early warning signs in potentially fraudulent companies, data on the pre-fraud period (early years) were compared to data on the fraud period (advanced years).
Post-Fraud Period

The post-fraud period consists of the year immediately following the fraud period. It is limited to one year as it was noticed that a number of fraud companies went bankrupt in one or two years following the fraud period and, for some companies, the audited annual reports were not available for several years following the discovery of fraud. For example, for fraud companies such as Waste Management Inc., Symbol Technologies Inc., and Thor Industries Inc., no audited annual reports were available for the post-fraud year. Data on the post-fraud year were collected to build the fraud model for detecting early warning signs of fraud and primarily used as a testing set for predicting the level of fraud.

Fiscal Year

The fiscal year for a company is an accounting period of 12 months that may be different from the calendar year. For example, for some companies, the fiscal year begins with April 1 and ends on March 31 or begins with July 1 and ends on June 30 or begins with November 1 and ends on October 31.

Fiscal years (company’s reporting period) were taken into consideration when defining the fraud period and collecting 10-Ks for the companies. For all fraud companies, each year of the fraud period matches the year of the company’s performance when it was alleged to be fraudulent. In most cases, the year of the 10-K matches the last year of the company’s performance except when a company fails to submit its 10-K within 90 days of the end of its fiscal year. In that case, the company issues a notification of late filing
(NT 10-K form) and files the 10-K when it is ready, which may be later than the year when it was required to file.

**Time Period**

In this study, the time period from 1993 to 2006 was selected for two reasons. First, 10-Ks are electronically available for download in the EDGAR database starting in 1994. EDGAR stores submissions by companies that file forms such as 10-Ks (annual reports) with the U.S. Securities and Exchange Commission (SEC). All domestic publicly-held companies with stock to trade are required to file their 10-Ks with the SEC. A second reason for selecting this time period is that a majority of high profile accounting scandals occurred during this period, which eventually led to the enactment of the Sarbanes-Oxley legislation in 2002. Due to the infrequency of financial statement fraud, this time period was selected because a relatively large number of fraud observations could be identified.

### 4.2 SAMPLE SELECTION AND DATA COLLECTION

A sample of 126 fraud companies with 405 fraud years identified in the alleged fraud period was selected from the years 1993 to 2006. Table 4.1 explains the sample selection for the fraud dataset. Initially, a comprehensive list of all those U.S. publicly listed companies where fraud had occurred and been discovered over the 14-year period from 1993 to 2006 was created. A total of 140 companies were identified during this time period. Out of these 140 companies, a total of 126 companies that had filed their 10-Ks with the SEC and whose 10-Ks were electronically available for download from the EDGAR online database were selected. The remaining 14 companies had stopped filing
10-Ks with the SEC due to delisting. Since the 10-Ks of these companies were not available for the fraud period, they were dropped from the fraud dataset.

**TABLE 4.1**

*Sample Selection*

<table>
<thead>
<tr>
<th>Description</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of fraud companies identified during the time period 1993 to 2006</td>
<td>140</td>
</tr>
<tr>
<td>(Sources: Lexis-Nexis, Compustat, <em>WSJ</em> Index, AAERs)</td>
<td></td>
</tr>
<tr>
<td>Less: Companies that stopped filing 10-Ks due to delisting</td>
<td>14</td>
</tr>
<tr>
<td>Final sample of companies selected</td>
<td>126</td>
</tr>
</tbody>
</table>

There were seven companies where multiple frauds were identified at different time periods. In such cases, multiple events were included in the fraud period for that company, but the company was counted only once in the sample. For example, for Bristol Myers Squibb, two frauds (one in each period) were identified for the periods 1999-2003 and 2006; as a result, both time periods were included in the fraud period, but the company was counted only once in the sample.

Figure 4.1 shows the yearly distribution of the fraud sample. It provides the number of observations by the last year of the fraud period. In cases of multiple frauds, the last year of the most recent fraud period was included. We can see from this chart that fraud companies almost doubled from 15 in 1998 to 28 in 2002. The number of fraud companies hit its peak in 2002 at 28 observations. In 2003, the number of fraud companies plummeted to 12 and in 2004 increased to 15. Since 2002, the annual rate has
fallen to about 10 companies per year. This cyclical pattern in fraudulent financial reporting suggests that this trend may continue. In spite of the anti-fraud measures taken by the government, such as enactment of Sarbanes-Oxley (SOX) in 2002, there are many instances of fraudulent financial reporting that continue to surface in the post-SOX period.

![Graph showing distribution of fraud companies by last year of fraud period]

**FIG. 4.1 – Distribution of Fraud Companies by Last Year of Fraud Period**

A total of 622 U.S. publicly listed corporations with 622 no-fraud years, where fraud had never been reported and whose 10-Ks are available for electronic download from the database for the time period 1993-2006, were also selected. As explained in the previous section, the selection of multiple no-fraud companies was done on the basis of year,
industry and size. An attempt was also made to select no-fraud firms proportional to the firms with occurrence of fraud during the fraud period. For example, 12 percent of the fraud companies (15 companies) were identified during the year 1998; therefore, approximately the same percent of no-fraud companies (75 companies) were selected from this period as well.

An effort was also made to ensure that the industry and sectors were also proportionately represented in the no-fraud dataset. For example, 11 fraud companies were identified in the ‘Utilities’ industry (two-digit SIC code 49); therefore, approximately 54 (about 8.7 percent) no-fraud companies were selected from this industry that met the criteria of size.

**TABLE 4.2**

*Distribution of Fraud and No-Fraud Companies by Industrial Sector*

<table>
<thead>
<tr>
<th>Industrial Sector (Industry Codes)</th>
<th>Fraud Companies</th>
<th>No-Fraud Companies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count</td>
<td>Percent</td>
</tr>
<tr>
<td>Mining (1000-1499)</td>
<td>2</td>
<td>1.6</td>
</tr>
<tr>
<td>Construction (1500-1799)</td>
<td>2</td>
<td>1.6</td>
</tr>
<tr>
<td>Manufacturing (2000-3999)</td>
<td>47</td>
<td>37.3</td>
</tr>
<tr>
<td>Transportation, Communication, Electric, Gas &amp; Sanitary Services (4000-4999)</td>
<td>17</td>
<td>13.5</td>
</tr>
<tr>
<td>Wholesale Trade (5000-5199)</td>
<td>6</td>
<td>4.8</td>
</tr>
<tr>
<td>Retail Trade (5200-5999)</td>
<td>5</td>
<td>3.9</td>
</tr>
<tr>
<td>Finance, Insurance &amp; Real Estate (6000-6799)</td>
<td>16</td>
<td>12.7</td>
</tr>
<tr>
<td>Services (7000-8999)</td>
<td>31</td>
<td>24.6</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>126</strong></td>
<td><strong>100.0</strong></td>
</tr>
</tbody>
</table>
Table 4.2 shows the distribution of fraud and no-fraud companies split by industrial sector. The 126 fraud companies were unevenly distributed across eight industrial sectors. With 47 fraud companies, the Manufacturing sector had the highest percentage (37.3 percent) of fraud companies. The services sector, with 31 fraud companies, was second at 24.6 percent of identified fraud companies. Among the eight sectors, the fraud companies were widely distributed across 31 two-digit SIC industries. Most prominent were 39 companies with SIC codes from 3000-3999 and 29 companies with SIC codes from 7000-7999. Most of these were technology companies.

For the 126 fraud companies, 10-Ks were also collected for pre-fraud years and post-fraud year ('fraud period-4,' 'fraud period-3,' 'fraud period-2,' 'fraud period-1,' 'fraud period,' 'fraud period+1'). Tables 4.3 and 4.4 explain both versions of the sample that were used for fraud detection. Table 4.5 explains the sample that was used for detecting early warning signs of fraud in potentially fraudulent companies. Since a fraud company may not have existed in certain years prior or subsequent to the fraud discovery, the number of fraud years varied due to missing data. The dataset in Table 4.5 was used to identify different levels (stages) of fraud in potentially fraudulent companies. Here, we seek to identify features that distinguish early warning signs of fraud from symptoms of advanced fraud in fraudulent companies. Since, the damage from fraud grows exponentially over time, it is important to detect it early. Persons (1995), Fanning and Cogger (1998), Lee et al. (1999), Kaminski et al. (2004), and Hoogs et al. (2007) also pointed out that early detection provides the best opportunity to minimize the damage that fraud can bring in potentially fraudulent companies.
### TABLE 4.3

*Sample for Fraud Detection - Version 1*

<table>
<thead>
<tr>
<th>Sample for fraud detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 405 fraud years (126 fraud companies)</td>
</tr>
<tr>
<td>• 622 no-fraud years (622 no-fraud companies)</td>
</tr>
</tbody>
</table>

### TABLE 4.4

*Sample for Fraud Detection - Version 2*

<table>
<thead>
<tr>
<th>Sample for fraud detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 405 fraud years (126 fraud companies)</td>
</tr>
<tr>
<td>• 622 no-fraud years (622 no-fraud companies)</td>
</tr>
<tr>
<td>• 348 no-fraud years (126 fraud companies)</td>
</tr>
</tbody>
</table>

### TABLE 4.5

*Sample for Detecting Early Warning Signs of Fraud*

<table>
<thead>
<tr>
<th>Sample for detecting early warning signs of fraud</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 208 pre-fraud years (early years of 126 fraud firms)</td>
</tr>
<tr>
<td>• 405 fraud years (advanced years of 126 fraud firms)</td>
</tr>
<tr>
<td>• 100 post-fraud years (126 fraud firms)</td>
</tr>
</tbody>
</table>
Moreover, for both sets of fraud and no-fraud companies, the original 10-Ks filed with SEC were collected and not the restated 10-Ks, also known as 10-K/A. The original 10-Ks were selected because a restatement of a financial statement is typically created to correct the previous financial statement for intentional/unintentional errors and accounting irregularities. Restatements represent an acknowledgement by the firm that prior financial statements were not in accordance with Generally Accepted Accounting Principles (Palmrose and Scholz, 2004). In order to identify symptoms of fraud and proactively detect fraud, we needed to examine and analyze the original 10-Ks and not the restated 10-Ks.

In addition, 8-K filings that were available for selected fraudulent and non-fraudulent companies during this time period were also collected. Form 8-K is the “current report” that companies must file with the SEC to let shareholders know about major events such as auditor changes, restatements, significant acquisitions, bankruptcy, large write-offs, major restructuring changes and so on. These episodic 8-K filings were collected as they sometimes contain information disclosures that are not otherwise reported and may provide additional insight. The dataset containing 8-K filings was not mixed with the dataset containing 10-Ks and was analyzed separately to gather additional insights about distinguishing features between fraudulent and non-fraudulent reports.

4.3 SUMMARY

This chapter presented the sample and data collection process. It also discussed the choice of data and rationale behind it. The sample consisted of 126 fraud and 622 no-fraud companies. A peer set for each fraud company was selected as opposed to a
matched-pair dataset. Both fraud and no-fraud companies were selected from 1993 to 2006. LexisNexis, Compustat, *WSJ* Index, and AAERs issued by the SEC were the primary sources of gathering all fraud company data. For both sets of companies, original 10-Ks were collected and not the restated 10-Ks. The 8-K filings that were available for the selected companies during this time period were also collected and examined separately from the dataset containing 10-Ks.
CHAPTER 5
BASELINE RESULTS

This chapter presents the results of the baseline approach for detecting fraud from the qualitative textual content of annual reports. Section 5.1 provides an overview of the baseline approach and the baseline features. Section 5.2 includes an overview of the classifiers used in this study. Section 5.3 presents baseline results for detecting fraud and for detecting early symptoms of fraud using a Naïve Bayes classifier. This section also includes top discriminative words listed by information gain scores. Section 5.4 discusses baseline results using a Support Vector Machine (SVM), the main classifier used in this study. This section also describes the creation of a document-term matrix for the initial feature set and subsequent reduction of its dimensionality without sacrificing any useful feature. The set of experiments in this chapter shows to what extent it is possible to detect fraud when no feature extraction or feature selection is performed. Finally, Section 5.5 summarizes the baseline results.

5.1 BASELINE APPROACH

For baseline experiments, we used a universally accepted technique for document classification, called “bag of words.” In a “bag of words” approach, a document is represented with a vector of word counts that appear in it. In this approach, the exact ordering of the words in a document is ignored; instead, the information on number of
occurrences of each word is retained as shown in Figure 5.1. It is also sometimes called a vector space model, where documents are tokenized and represented as vectors in a multi-dimensional Euclidean space, each dimension in this space corresponding to a term (token). The value for each dimension is the frequency of word (token) occurrence in the document. Another variation of the “bag of words” approach is called “bag of bits.” In this approach, a document is modeled as a vector of bits, i.e., a vector of 0’s and 1’s. In a “bag of bits” approach, if a word appears in a document at least once, it is counted as 1, otherwise 0.

Using a “bag of words” approach, we ran baseline experiments, where the documents (10-Ks) are represented by a vector of its word counts. Here, the term frequency is used as the weight for each feature. For initial baseline experiments, preliminary data preprocessing was done in three steps. First, all the words were converted into lower case so that no two same words such as ‘allege’ and ‘Allege’ are included in the corpus as different words. Second, punctuation was removed. Third, numbers were removed from all the documents. Since, numbers contained in 10-Ks by themselves do not carry discriminative information, the input to the classifiers consisted of only non-numeric content. Each element of the feature vector space represented a unique word that occurs in the whole corpus and each non-zero value in a document (instance) indicated the number of times a word appeared in that document. This resulted in a very high-dimensional feature space due to large number of features (words).

At this stage, we did not use a stop words list nor did we perform stemming, as we did not want to throw away any useful token. We evaluated the stop words separately and adjusted the stop words list so that it did not include any of those words that are relevant
for our study. For instance, auxiliary verbs were not included in the stop words list, as these tokens were required to analyze uncertainty marker features. The results of the baseline experiments without applying a stop words list and with an adjusted stop words list are explained in the next two sections. As discussed earlier under ‘Preprocessing Dataset’ (Chapter 3), stemming was not performed as inflected variants of base forms often have very different meanings in the domain of accounting (see, for example, Chen et al., 1995; Garnsey, 2006).

In a “bag of words” approach, the learning algorithm examines the bag of words vector associated with the incoming document and sees if it fits closely to typical vectors associated with a given class or not. Two documents with similar “bag of words” representations are considered similar in content. Figure 5.1 illustrates the “bag of words” approach with an example.
(a) Snippet from an Annual Report (10-K) of Company A

The Company is a defendant in a number of other pending legal proceedings incidental to present and former operations, acquisitions and dispositions. The Company does not expect the outcome of these proceedings, either individually or in the aggregate, to have a material adverse effect on its financial position, results of operations or liquidity.

(b) Snippet from an Annual Report (10-K) of Company B

In the opinion of the Company, the outcome of any of these matters will not have a material adverse effect on the Company's consolidated financial position but could be material to its results of operations in any one accounting period.

(c) “Bag of Words” for the 10-K snippets shown in (a) & (b)

\( w_1 = \text{the, } w_2 = \text{company, } w_3 = \text{is, } w_4 = \text{a, } w_5 = \text{defendant, } w_6 = \text{in, } w_7 = \text{number, } w_8 = \text{of, } w_9 = \text{other, } w_{10} = \text{pending, } w_{11} = \text{legal, } w_{12} = \text{proceedings, } w_{13} = \text{incidental, } w_{14} = \text{to, } w_{15} = \text{present, } w_{16} = \text{and, } w_{17} = \text{former, } w_{18} = \text{operations, } w_{19} = \text{acquisitions, } w_{20} = \text{dispositions, } w_{21} = \text{does, } w_{22} = \text{not, } w_{23} = \text{expect, } w_{24} = \text{outcome, } w_{25} = \text{these, } w_{26} = \text{either, } w_{27} = \text{individually, } w_{28} = \text{or, } w_{29} = \text{aggregate, } w_{30} = \text{have, } w_{31} = \text{material, } w_{32} = \text{adverse, } w_{33} = \text{effect, } w_{34} = \text{on, } w_{35} = \text{its, } w_{36} = \text{financial, } w_{37} = \text{position, } w_{38} = \text{results, } w_{39} = \text{liquidity, } w_{40} = \text{opinion, } w_{41} = \text{any, } w_{42} = \text{matters, } w_{43} = \text{will, } w_{44} = \text{companies, } w_{45} = \text{consolidated, } w_{46} = \text{but, } w_{47} = \text{could, } w_{48} = \text{be, } w_{49} = \text{one, } w_{50} = \text{accounting, } w_{51} = \text{period} \)

(d) Document Representation Version of the Partial 10-Ks shown in (a) & (b)

\( A = (4, 2, 1, 2, 1, 3, 1, 1, 1, 2, 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) \)

\( B = (4, 1, 0, 1, 0, 2, 0, 4, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 2, 1, 1, 1, 1, 1, 1, 0, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1) \)

**FIG. 5.1 – Illustration of “Bag of Words” Approach**

88
5.2 METHODS

In this section we provide an overview of the classifiers that we used in this study. We first discuss the Naïve Bayes classifier that we used to run our preliminary baseline experiments and then we discuss Support Vector Machine, the main classifier used in this study. This discussion of the tools is followed by the presentation and discussion of baseline results that we obtained with each classifier.

5.2.1 NAÏVE BAYES

Naïve Bayes (NB) is one of the simplest and most effective inductive learning algorithms. The basic idea in the NB approach is to use the joint probabilities of words and categories to estimate the probabilities of categories when a document is given (McCallum and Nigam, 1998). The NB classifier assigns the most likely class to a given example described by its feature vector. The underlying assumption of the NB approach is that the probability of each word occurring in a document is independent of the occurrence of other words in the document and the probability that a document is generated in some class depends only on the probabilities of the words given the context of the class. Even though it is a probabilistic classifier, its classification performance is competitive with the performance of other sophisticated machine learning methods (Mitchell, 1997).

In text categorization, two types of models have been used: the multi-variate Bernoulli model and the multinomial model (McCallum and Nigam, 1998). One of the main differences between the two models is that the multi-variate Bernoulli model captures the information on which words are present or absent in a document, but not the number of
times each word occurs in a document whereas multinomial model also captures the information about how many times a word occurs in a document. McCallum and Nigam (1998) have shown that for larger vocabulary sizes the multinomial model outperforms the multi-variate model whereas for smaller fixed vocabulary sizes the multi-variate model performs reasonably well.

Rish (2001) conducted an empirical study of the NB classifier and explains that a classifier is defined by a deterministic function that assigns a class to any given example (instance). A typical approach is to associate each class with a discriminant function and let the classifier select the class with the maximum discriminant function on a given example. The NB classifier uses the class posterior probability given a feature vector as its discriminant function and simply selects the class with the highest posterior probability.

Under the NB approach, the task of assigning a new unlabelled document $D$ with feature vector $W = (w_1, w_2, ..., w_n)$ to a class $c$ where $c$ can take one of $m$ values $c \in \{0, 1, ..., m\}$ can be defined by a function $f_{NB}^*$ as:

$$f_{NB}^* = \arg \max_{c_j \in C} \prod_{i=1}^{n} P(w_i | c_j)$$

(5.1)

where $P(c_j)$ is the prior probability of class $c_j$, and

$$P(w_i | c_j)$$

is the posterior probability of word $w_i$ given class $c_j$.

$P(c_j)$ and $P(w_i | c_j)$ can be computed based on their frequency in the training data. This is explained for a binary classification task. In this case, we have two classes +1 and -1. A
feature vector consisting of n features (words) describes each document in the corpus. The value of each feature (word) corresponds to the number of times it occurs in a given document. \( P(c_j) \) is calculated by counting the number of instances with \( c = +1 \) and with \( c = -1 \). \( P(w_i | c_j) \) is estimated by solving for two sub-problems: \( P(w_i | c_{j+1}) \) and \( P(w_i | c_{j-1}) \). Due to the independence assumption made by NB, \( P(w_i | c_j) \) can be written as \( P(w_i | c_j) \) \( P(w_2 | c_j) \ldots P(w_n | c_j) \). This implies that \( P(w_i | c_{j+1}) \) and \( P(w_i | c_{j-1}) \) become \( P(w_i | c_{j+1}) \) \( P(w_2 | c_{j+1}) \ldots P(w_n | c_{j+1}) \) and \( P(w_i | c_{j-1}) \) \( P(w_2 | c_{j-1}) \ldots P(w_n | c_{j-1}) \) respectively.

The formulation given in Equation 5.1 can be problematic in cases of zero probabilities. This takes place when a given class and feature never occur together in the training set. This causes \( P(w_i | c_j) \) to amount to zero when the individual probabilities of the attributes over the feature vector are multiplied even if the test document contains other terms. This problem of zeros is handled by incorporating a Laplace correction to smooth word probabilities or by assigning the probability 1/N to the feature, where N is the number of examples in the training set. The Laplace correction estimator assumes that the observation of each word occurring is a priori equally likely. Joachims (1997) found that this Bayesian Laplace estimator works well in practice, as it does not falsely estimate probabilities to be zero.

Even though the assumption of independence in NB is unrealistic for real world examples, its performance has been found to be superior to other learning algorithms, even in datasets with substantial feature dependencies. Domingos and Pazzani (1997) attribute the good performance of NB to the ‘zero-one’ loss function. This function
defines error equal to the number of incorrect classifications. It should be noted that this function does not penalize inaccurate probability estimates as long as the maximum probability is assigned to the correct class. Thus, NB may revise the posterior probability of each class, but its classification accuracy is not affected as the class with maximum posterior probability rarely changes.

**5.2.2 SUPPORT VECTOR MACHINES**

Support Vector Machine (SVM) is a statistical supervised machine learning technique that has been found to be particularly useful for classification tasks such as text classification and document classification. It is based on the Structural Risk Minimization (SRM) principle drawn from statistical learning theory (Vapnik, 1982). The objective of SRM is to find a hypothesis h from a hypothesis space H that guarantees the lowest probability of error E(h), i.e., the lowest probability of making a false prediction on an example that is randomly drawn from the sample for a given training sample S consisting of n examples. The fundamental assumption underlying all existing classification methods, both linear and nonlinear, is that all training data can be classified by a single decision boundary, no matter how complex the classifier is (Burges, 1998). Under SVM, the basic idea is to achieve the well-known trade-off between the complexity of the hypothesis space and the training error. A simple hypothesis may generate high training error whereas a complex hypothesis may lead to a small training error at the cost of overfitting the data. Thus, it is critical to choose the hypothesis that is of right complexity.
In the case of a linearly separable binary classification problem, the two classes represented by training examples (positive, negative) are perfectly separable by a separating hyperplane as seen in Figure 5.2. This figure presents a graphical representation of a linear SVM in a simple two-dimensional input space. Each feature corresponds to one dimension in this feature space. All vectors lying on one side of the separating hyperplane are labeled as ‘+1’ and belong to the positive class, and all vectors lying on the other side of this hyperplane are labeled as ‘-1’ and belong to the negative class. In the case of fraud detection, the two classes—fraud and no-fraud—are assigned the numerical class label ‘+1’ and ‘-1,’ respectively. The distance from the separating hyperplane to a data point is determined by the strength of each feature of the data. Thus, in the case of fraud classification, an annual report belonging to the positive (fraud) class with many strong features related to fraud would be located far from the class boundary on the positive side whereas an annual report belonging to negative (no-fraud) class with
no features related to fraud but having many no-fraud related features would be located far from the class boundary on the negative side.

The training data $D$, containing $n$ labeled examples, is given by:

$$D = \{(x_i, y_i), i = 1, \ldots, n, x_i \in \mathbb{R}^m, y_i \in \{-1, +1\}\}$$ (5.2)

where each $x_i$ is a vector of features describing example $i$ and each $y_i$ is the class label for that example; $x_i$ are vectors in $m$-dimensional input space $\mathbb{R}^m$; size $n$ is equal to the number of examples in the given training data.

In order to decide which of the two class predictions is correct for an unseen data point, a linear SVM uses a hypothesis vector $w$ and bias term $b$ to classify a new example $x$, by generating a predicted class label $f(x)$.

The classification (decision) function for an unlabeled document $x$, using a SVM with separating hyperplane $(w, b)$ is given by:

$$f(x) = \text{sign} \left( <w, x> + b \right)$$ (5.3)

where $w$ is the vector of learned weights and $x$ is an input vector (representing a new unlabeled document); $b$ (bias) denotes the perpendicular distance from the hyperplane to the origin; the separating hyperplane is defined by the equation:

$$<w, x> + b = 0 \text{ where } w \in \mathbb{R}^m, b \in \mathbb{R}$$ (5.4)

As a result, SVM learning for linearly separable data (maximizing margin) can be viewed as the following optimization problem:
minimize $\frac{||w||^2}{2}$ \quad \text{subject to: } y_i (< w, x_i > + b) \geq 1; \quad \forall i = 1, \ldots, n \quad (5.5)$

where $x_i$ is a training example with label $y_i$ and $||w||$ is the Euclidean length of the weight vector $w$ (the Euclidean norm of $w$); $n$ denotes the number of training examples; vector $w$ and bias $b$ define the position and orientation of the hyperplane.

The optimization problem defined in Equation 5.5 has a dual formulation. The dual solution is useful as it allows for higher dimensional feature space by collapsing the data points into a matrix of inner products and allows for generalization to the nonlinear case.

The corresponding dual problem is as follows:

$\text{maximize } \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) \quad \text{subject to: } \sum_{i=1}^{n} y_i \alpha_i = 0; \alpha_i \geq 0; \quad i = 1, \ldots, n \quad (5.6)$

where $\alpha_i$ and $\alpha_j$ are the Lagrange multipliers.

In the case of dual formulation, the decision function defined in Equation 5.3 becomes:

$f(x) = \text{sign} \left( \sum_{i=1}^{n} y_i \alpha_i x_i \cdot x + b \right) \quad (5.7)$

The distance of the hyperplane to the nearest of the positive and negative examples is known as the margin of the SVM. For the linearly separable case, the support vector algorithm simply looks for the hyperplane that separates the two classes of the data with the largest margin. There are many different hyperplanes that can perform separation, but
the objective of SVM is to find the hyperplane with the widest margin out of all the hyperplanes that minimize the upper bound on the generalization error. In order to calculate the margin, SVM constructs two parallel hyperplanes, one on each side of the separating hyperplane and chooses the one with the widest margin, as seen in Figure 5.3.

The examples that are closest to the hyperplane or lie on the hyperplane are called support vectors. In general, a classifier with smaller margins will have a higher expected risk when facing previously unseen data; the wider the margin, the better is the generalization error of the classifier. In other words, classifiers having smaller margins tend to have poorer generalizability than classifiers with wider margins. The main assumption in the case of a perfect classification is that the majority of the data can be described by a single linear decision boundary and data points that do not fit this linear decision boundary are called outliers.

FIG. 5.3 – (a) Hyperplane with Smallest Margin, (b) Hyperplane with Widest Margin
Source: Cristianini & Shawe-Taylor (2000)
If classes are linearly inseparable and it is not feasible to achieve linear decision hyperplanes, then the input space is mapped to a higher dimensional feature space and true separation is achieved by a quadratic boundary. When we use a nonlinear separation boundary, we are able to separate two classes without any error and there are no misclassifications, as shown in Figure 5.4. It should be noted that the true separation is not achieved by a linear line but rather by a nonlinear line, which is a quadratic curve. With a linear separation, there were four misclassified negative cases and two misclassified positive cases. However, all of these misclassified cases were correctly separated when a nonlinear (complex) decision boundary was used.

FIG. 5.4 – Case of Nonlinear Classification

Source: Cristianini & Shawe-Taylor (2000)
For the case of nonlinear classification, two alternative formulations of linear optimization problem are proposed: one is based upon slack variables and the other is based upon using kernels. The optimization problem for SVM with kernels transforms into:

$$\text{maximize} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \quad \text{subject to}: \sum_{i=1}^{n} y_i \alpha_i = 0, \alpha_i \geq 0; \ i = 1, \ldots, n$$  \hspace{1cm} (5.8)

with the decision rule as follows:

$$f(x) = \text{sign} \left( \sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + b \right)$$  \hspace{1cm} (5.9)

where $K$ is a kernel function.

Furthermore, in case of a nonlinear classification problem, training data (input vectors) can only appear in the form of scalar (inner) products $(x_i, x_j)$. The mapping $\Phi$ is used to transform data from the input space to the feature space and the function that calculates the inner product between mapped examples in a feature space is called a kernel function. Thus, products in input space are mapped to some other dot product $\Phi(x_i) \cdot \Phi(x_j)$ in feature space $F$ via a nonlinear mapping $\Phi: \mathbb{R}^m \to F$. In this feature space, the required scalar products are calculated directly by computing kernels $K(x_i, x_j)$ for the given training data vectors in an input space and are expressed by using a kernel function $K(x_i, x_j) = <\Phi(x_i) \cdot \Phi(x_j)>$, which is a function in input space. It should be noted that the kernel computes the inner product by implicitly mapping the examples to the high dimensional
feature space thus allowing for more unique features without increasing computational cost. The hyperplanes in this high dimensional feature space correspond to more complex decision boundaries in the original input space. This way, by applying the kernel trick (i.e., as long as $K(x_i, x_j)$ can be calculated efficiently, there is actually no need to explicitly map the data into feature space using $\Phi$), SVMs handle a nonlinear classification efficiently by avoiding both overfitting and underfitting of the training data.

To briefly recapitulate, SVM consists of three main phases: (1) input or transformation phase, (2) learning phase, and (3) decision phase. In the input phase, SVM does its most significant work of transforming data into a high dimensional feature space by using kernel mapping. In the learning phase, SVM starts to learn the data from the high dimensional feature space and takes into consideration the dot product of the feature vectors in addition to minimizing the magnitude of the weight vectors to construct the optimal hyperplane by extracting support vectors only. In the decision phase, SVM produces the final output function using these support vectors’ information. When constructing the final decision function, SVM does not consider all samples but still manages to obtain the unique solution for the decision function. In addition, Vapnik (1995) has shown that unlike most traditional learning algorithms, SVM minimizes the structural risk (which minimizes the upper bound on the generalization error) rather than the empirical risk (which minimizes the error on the training data). Another important advantage of SVMs over conventional classifiers is that they do not need any parameter tuning as they can find good parameter settings automatically.
5.2.2.1 SEQUENTIAL MINIMAL OPTIMIZATION ALGORITHM

Waikato Environment for Knowledge Analysis (WEKA) uses the Sequential Minimal Optimization (SMO) algorithm to perform the training of SVMs. In this section, we will briefly discuss how the SMO algorithm works. The SMO algorithm breaks down the large quadratic programming optimization problem into a series of smallest possible quadratic problems. These small quadratic problems consist of a minimal subset of just two data points and are solved analytically without the need to use an iterative quadratic program optimizer as part of the algorithm (Platt, 1998). At each iteration, SMO chooses only two Lagrange multipliers $\alpha_i$ to jointly optimize them, while keeping other multipliers constant. It then finds the optimal values for these two parameters, and updates the SVM to reflect the new optimal values. The two components of SMO are: an analytic method for solving for the two Lagrange multipliers, and a heuristic for choosing the multipliers to optimize. The SMO algorithm uses a heuristic motivated by the Karush-Kuhn-Tucker (KKT) conditions to choose the two multipliers and performs the optimization of these two parameters analytically. At every step, both Lagrange multipliers are replaced with new multipliers that are chosen via good heuristics (KKT).

It basically generates an initial vector set and, at each step, adds a single training vector to the set of existing support vectors and eliminates support vectors that are linearly dependent in the feature space. Thus, every time a new vector is added to the existing set, another vector is removed and in this process it eliminates many unnecessary features also with those vectors. This continues until the algorithm converges to the optimal solution. The goal of the SMO algorithm is to optimize the global problem by acting on a small subset of data at a time (Platt, 1999). The decomposition method allows the SMO
algorithm to handle very large training sets using less computation time. In addition, the SMO algorithm has been found to be particularly fast for linear SVMs and sparse datasets, which are often encountered in text classification tasks.

### 5.2.2.2 LINEAR KERNELS

There are many types of kernels that are used to train support vector machines such as linear, polynomial, sigmoid, and Gaussian radial basis function kernels as shown in Table 5.1. The simplest kernel is a linear kernel defined as $K(x_i, x_j) = x_i^T x_j$. The choice of the kernel function plays a critical role in the efficiency of SVMs. Different kernels have been shown to yield the best performance depending upon the learning task at hand.

For our dataset, we constructed a linear classifier, which determines the separating hyperplane that has maximum distance from the closest points of the training set, also called margins. As explained earlier, linear kernels have been used efficiently with very large training sets, both in terms of the number of examples and the dimensionality of input space. In general, linear kernels are used when the number of features is large and the number of instances is less than the number of features.

We used simple linear SVMs in the experiments because they are faster, efficient, simple to use, perform well on high-dimensional problems and provide high generalization accuracy with default parameter settings. In addition, one of the main advantages of using linear kernels is that there is less danger of overfitting as the more complex a kernel is, the more chances there are that it would suffer from the problem of overfitting by trying to fit more training data, which degrades the classifier accuracy. Existing literature on
text categorization also indicates that nonlinear SVMs – polynomial and radial basis functions gain very little in terms of performance over linear models (Joachims, 1998).

**TABLE 5.1**

*Different Kernels Available to Train SVMs*

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>$K(x_i, x_j) = x_i^T x_j$</td>
</tr>
<tr>
<td>Polynomial</td>
<td>$K(x_i, x_j) = (x_i^T x_j + 1)^d$, where $d$ is the degree of the polynomial</td>
</tr>
<tr>
<td>Radial Basis Function (RBF)</td>
<td>$K(x_i, x_j) = \exp(-\gamma |x_i-x_j|^2)$, where $\gamma$ is the width of the RBF function and $\gamma &gt; 0$</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + 1)$, $\gamma &gt; 0$</td>
</tr>
</tbody>
</table>

**5.3 BASELINE RESULTS USING NAÏVE BAYES CLASSIFIER**

For our initial baseline results, we used the Bow (also known as Rainbow) classifier system based on the Bow library that uses the Naïve Bayes (NB) algorithm as the default algorithm for text classification. Rainbow is a statistical modeling toolkit for text classification that was developed by Andrew McCallum. Rainbow has options for both models of Naïve Bayes namely the multi-variate Bernoulli model and the multinomial model. For preliminary experiments we use the multinomial model, which has been shown to perform well with large feature sets (McCallum and Nigam, 1998). Even though Rainbow is primarily designed for classification by the NB algorithm, it supports various other classification methods such as TFIDF/Rocchio, Probabilistic Indexing and K-nearest neighbor. Bow library is written in C and compiles on most UNIX systems.
The reasons for using the Naïve Bayes classifier for our initial baseline results were: (1) it is easy to implement, (2) it is among the most successful known algorithms after SVM for text classification (Dumais et al., 1998), (3) the experimental results helped us compare the performance of the two popular text classifiers (NB and SVM) and back our claims that SVM is better suited to our problem of fraud detection, and (4) it acted as a preprocessor to explore useful feature subsets for SVM.

**5.3.1 EXPERIMENTAL RESULTS AND DISCUSSION**

We applied NB classification to the problem of document categorization in which we focused on two issues: (1) fraud detection, and (2) detection of different stages (levels) of fraud. We studied each of these two issues on two versions of datasets. The data used for our experiments consisted of the 10-Ks downloaded from EDGAR database (see Chapter 4). The first version of the dataset for detecting fraud consisted of 1027 documents belonging to two categories (fraud, no-fraud) for training and testing datasets. Out of these 1027 documents, 405 documents were fraudulent 10-Ks of 126 fraud companies and 622 documents were non-fraudulent 10-Ks of 622 no-fraud companies. The second version of the dataset for detecting fraud consisted of 1375 documents belonging to two categories (fraud, no-fraud) for training and testing datasets. Out of these 1375 documents, 405 documents were fraudulent 10-Ks of 126 fraud companies and the remaining 970 documents consisted of 622 non-fraudulent 10-Ks of 622 no-fraud companies and 348 non-fraudulent 10-Ks of 126 fraud companies, which were outside the pre-fraud, fraud and post-fraud period (see Chapter 4 for more information on sample selection and data collection).
The first version of the dataset for detecting different stages of fraud consisted of 713 documents belonging to three categories (pre-fraud, adv-fraud, post-fraud) for training and testing datasets. Out of these 713 documents, 208 documents were 10-Ks of pre-fraud years of 126 fraud companies, 405 documents were 10-Ks of adv-fraud years of 126 fraud companies, and 100 documents were 10-Ks of post-fraud years of 126 fraud companies. The second version of the dataset for detecting different stages of fraud consisted of 613 documents belonging to two categories (pre-fraud, adv-fraud) for training and testing datasets. Out of these 613 documents, 208 documents were 10-Ks of pre-fraud years of 126 fraud companies, and 405 documents were 10-Ks of adv-fraud years of 126 fraud companies.

### TABLE 5.2

*Training and Testing Datasets*

<table>
<thead>
<tr>
<th></th>
<th>Total # of documents</th>
<th>Training data</th>
<th>Testing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraud Detection (Version 1)</td>
<td>1027</td>
<td>925</td>
<td>102</td>
</tr>
<tr>
<td>Fraud Detection (Version 2)</td>
<td>1375</td>
<td>1238</td>
<td>137</td>
</tr>
<tr>
<td>Detection of Stages of Fraud (Version 1)</td>
<td>713</td>
<td>642</td>
<td>71</td>
</tr>
<tr>
<td>Detection of Stages of Fraud (Version 2)</td>
<td>613</td>
<td>552</td>
<td>61</td>
</tr>
</tbody>
</table>

All 10-Ks were saved as text files, one file per document. Before performing classification or any other kind of diagnostics, all the documents (10-Ks) were submitted
to Rainbow so that it could index\(^5\) the data and create a model containing their statistics, which was then used for classification. The NB classifier was trained using 10-fold cross-validation. Table 5.2 shows the split between training and testing datasets.

### TABLE 5.3

*Baseline Results with NB Classifier*

<table>
<thead>
<tr>
<th>Features</th>
<th>Datasets</th>
<th>Fraud Detection</th>
<th>Detection of Stages of Fraud</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Fraud/NoFraud</td>
<td>Pre/Adv/Post</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Version 1</td>
<td>Version 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fraud/NoFraud</td>
<td>Pre/Adv/Post</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Version 2</td>
<td>Version 2</td>
</tr>
<tr>
<td>Bag of Words (TF)</td>
<td>(w/o applying stop words and no pruning)</td>
<td>52.17 %</td>
<td>42.54 %</td>
</tr>
<tr>
<td>Bag of Words (TF)</td>
<td>(with stop words only)</td>
<td>55.28 %</td>
<td>40.28 %</td>
</tr>
<tr>
<td>Bag of Words (TF)</td>
<td>(with pruning only)</td>
<td>55.11 %</td>
<td>41.97 %</td>
</tr>
<tr>
<td>Bag of Words (TF)</td>
<td>(with pruning and stop words)</td>
<td>56.75 %</td>
<td>39.01 %</td>
</tr>
</tbody>
</table>

We obtained baseline results by using simple feature reduction techniques (stop words, pruning) with a “bag of words” approach to explore their potential to improve classification accuracy. Table 5.3 shows a comparison of the average classification accuracy rates using 10-fold cross-validation for fraud datasets. These results indicate that for the first dataset, NB performed best when we applied both pruning and stop words. For the second dataset, NB performed best when we applied pruning only. For the

---

\(^5\) In the context of classification, indexing refers to the task of automatically constructing an internal representation of the data that can be interpreted by the classifier induction algorithm.
third dataset NB performed best when we applied neither pruning nor stop words, and its performance improved since the number of features increased. For the fourth dataset, NB performed best when we applied pruning only.

For detecting different levels (stages) of fraud, our initial baseline results showed that when the classifier was trained and tested on data of three classes (pre-fraud, adv-fraud, post-fraud), its best performance score was 42.54 percent. On the other hand, when we trained and tested the classifier with data of only two classes (pre-fraud, adv-fraud), its performance increased to 54.58 percent. The analysis of classifier errors indicated that the classifier misclassified all the instances of post-fraud class by assigning 90 percent of the instances to the adv-fraud class and 10 percent of the instances to the pre-fraud class.

These results suggest that most instances of post-fraud class exhibit strong symptoms of the adv-fraud class, and the annual reports for the year immediately following the time period of alleged fraud do not appear to be much different from the annual reports issued during the alleged fraud period. Most of the misclassifications seem to occur due to the large overlap of terms between the annual reports issued during the fraud period and the post-fraud period.

Another likely reason for poor performance of the classifier was that the training dataset in the minority categories such as post-fraud was too small to provide adequate training data. This was our motivation to collapse categories in the second version of the dataset relating to detection of stages of fraud. We carried out the rest of our experiments for this dataset with two categories only—pre-fraud and adv-fraud.
5.3.2 INFORMATION GAIN

Information Gain (IG) is the reduction of entropy with respect to the classification of a target class based on the observation of a feature. In other words, IG indicates how useful a feature is in predicting a class. The basic idea of IG is to retain features that reveal the most information about the distribution of classes. A text feature selection algorithm typically retains words with higher scores and discards words with smaller scores as words with smaller scores are rarely informative and do not contribute much in the prediction of the class. Very often, features whose IG score is less than some predetermined threshold are removed. We use an information gain measure to explore the discriminative power of each unique term and rank the features by the IG score. IG can be computed by subtracting the conditional entropy of the class from the total entropy of the class. Tables 5.4 and 5.5 list the top 78 and 73 discriminative words by information gain for detecting fraud and levels of fraud, respectively.
TABLE 5.4

*Ranking of the Top 78 Discriminative Words by Information Gain for Detecting Fraud*

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>Rank</th>
<th>Word</th>
<th>Rank</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>allege</td>
<td>27</td>
<td>less</td>
<td>53</td>
<td>latin</td>
</tr>
<tr>
<td>2</td>
<td>argentina</td>
<td>28</td>
<td>singapore</td>
<td>54</td>
<td>employees</td>
</tr>
<tr>
<td>3</td>
<td>brazil</td>
<td>29</td>
<td>notional</td>
<td>55</td>
<td>tax</td>
</tr>
<tr>
<td>4</td>
<td>plaintiffs</td>
<td>30</td>
<td>hedged</td>
<td>56</td>
<td>whether</td>
</tr>
<tr>
<td>5</td>
<td>alleges</td>
<td>31</td>
<td>provided</td>
<td>57</td>
<td>equity</td>
</tr>
<tr>
<td>6</td>
<td>defendants</td>
<td>32</td>
<td>interest</td>
<td>58</td>
<td>germany</td>
</tr>
<tr>
<td>7</td>
<td>manhattan</td>
<td>33</td>
<td>past</td>
<td>59</td>
<td>barbados</td>
</tr>
<tr>
<td>8</td>
<td>cooperating</td>
<td>34</td>
<td>days</td>
<td>60</td>
<td>obtain</td>
</tr>
<tr>
<td>9</td>
<td>purported</td>
<td>35</td>
<td>over</td>
<td>61</td>
<td>beginning</td>
</tr>
<tr>
<td>10</td>
<td>venezuela</td>
<td>36</td>
<td>market</td>
<td>62</td>
<td>payable</td>
</tr>
<tr>
<td>11</td>
<td>use</td>
<td>37</td>
<td>amounts</td>
<td>63</td>
<td>policies</td>
</tr>
<tr>
<td>12</td>
<td>none</td>
<td>38</td>
<td>share</td>
<td>64</td>
<td>flows</td>
</tr>
<tr>
<td>13</td>
<td>held</td>
<td>39</td>
<td>value</td>
<td>65</td>
<td>purchase</td>
</tr>
<tr>
<td>14</td>
<td>paid</td>
<td>40</td>
<td>stock</td>
<td>66</td>
<td>holders</td>
</tr>
<tr>
<td>15</td>
<td>about</td>
<td>41</td>
<td>vigorously</td>
<td>67</td>
<td>standards</td>
</tr>
<tr>
<td>16</td>
<td>aggregate</td>
<td>42</td>
<td>putative</td>
<td>68</td>
<td>security</td>
</tr>
<tr>
<td>17</td>
<td>outstanding</td>
<td>43</td>
<td>islands</td>
<td>69</td>
<td>term</td>
</tr>
<tr>
<td>18</td>
<td>price</td>
<td>44</td>
<td>opinion</td>
<td>70</td>
<td>unspecified</td>
</tr>
<tr>
<td>19</td>
<td>taxes</td>
<td>45</td>
<td>cash</td>
<td>71</td>
<td>philippines</td>
</tr>
<tr>
<td>20</td>
<td>counterparties</td>
<td>46</td>
<td>notes</td>
<td>72</td>
<td>counterparty</td>
</tr>
<tr>
<td>21</td>
<td>colombia</td>
<td>47</td>
<td>capital</td>
<td>73</td>
<td>costs</td>
</tr>
<tr>
<td>22</td>
<td>seeks</td>
<td>48</td>
<td>total</td>
<td>74</td>
<td>transactions</td>
</tr>
<tr>
<td>23</td>
<td>shares</td>
<td>49</td>
<td>common</td>
<td>75</td>
<td>austria</td>
</tr>
<tr>
<td>24</td>
<td>plan</td>
<td>50</td>
<td>their</td>
<td>76</td>
<td>allegations</td>
</tr>
<tr>
<td>25</td>
<td>requirements</td>
<td>51</td>
<td>present</td>
<td>77</td>
<td>complaints</td>
</tr>
<tr>
<td>26</td>
<td>generally</td>
<td>52</td>
<td>believe</td>
<td>78</td>
<td>assurance</td>
</tr>
</tbody>
</table>
### TABLE 5.5

*Ranking of the Top 73 Discriminative Words by Information Gain for Predicting Levels of Fraud*

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>Rank</th>
<th>Word</th>
<th>Rank</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sarbanes</td>
<td>26</td>
<td>breached</td>
<td>51</td>
<td>unvested</td>
</tr>
<tr>
<td>2</td>
<td>weaknesses</td>
<td>27</td>
<td>complies</td>
<td>52</td>
<td>filings</td>
</tr>
<tr>
<td>3</td>
<td>oxley</td>
<td>28</td>
<td>misstatements</td>
<td>53</td>
<td>us</td>
</tr>
<tr>
<td>4</td>
<td>sros</td>
<td>29</td>
<td>guidance</td>
<td>54</td>
<td>unspecified</td>
</tr>
<tr>
<td>5</td>
<td>conclusions</td>
<td>30</td>
<td>deficiencies</td>
<td>55</td>
<td>persuasive</td>
</tr>
<tr>
<td>6</td>
<td>misleading</td>
<td>31</td>
<td>task</td>
<td>56</td>
<td>likelihood</td>
</tr>
<tr>
<td>7</td>
<td>qualitative</td>
<td>32</td>
<td>governance</td>
<td>57</td>
<td>my</td>
</tr>
<tr>
<td>8</td>
<td>omit</td>
<td>33</td>
<td>recoverability</td>
<td>58</td>
<td>outcome</td>
</tr>
<tr>
<td>9</td>
<td>eitf</td>
<td>34</td>
<td>alleges</td>
<td>59</td>
<td>looking</td>
</tr>
<tr>
<td>10</td>
<td>untrue</td>
<td>35</td>
<td>impairments</td>
<td>60</td>
<td>deterioration</td>
</tr>
<tr>
<td>11</td>
<td>quantitative</td>
<td>36</td>
<td>plaintiffs</td>
<td>61</td>
<td>weakness</td>
</tr>
<tr>
<td>12</td>
<td>fraud</td>
<td>37</td>
<td>forecasted</td>
<td>62</td>
<td>scholes</td>
</tr>
<tr>
<td>13</td>
<td>certifying</td>
<td>38</td>
<td>exit</td>
<td>63</td>
<td>motions</td>
</tr>
<tr>
<td>14</td>
<td>fasb</td>
<td>39</td>
<td>projections</td>
<td>64</td>
<td>putative</td>
</tr>
<tr>
<td>15</td>
<td>summarize</td>
<td>40</td>
<td>evaluated</td>
<td>65</td>
<td>certifications</td>
</tr>
<tr>
<td>16</td>
<td>llc</td>
<td>41</td>
<td>allegations</td>
<td>66</td>
<td>gov</td>
</tr>
<tr>
<td>17</td>
<td>impaired</td>
<td>42</td>
<td>deteriorate</td>
<td>67</td>
<td>corrections</td>
</tr>
<tr>
<td>18</td>
<td>com</td>
<td>43</td>
<td>defend</td>
<td>68</td>
<td>certification</td>
</tr>
<tr>
<td>19</td>
<td>defendants</td>
<td>44</td>
<td>bulletin</td>
<td>69</td>
<td>web</td>
</tr>
<tr>
<td>20</td>
<td>allege</td>
<td>45</td>
<td>complaints</td>
<td>70</td>
<td>criteria</td>
</tr>
<tr>
<td>21</td>
<td>ethics</td>
<td>46</td>
<td>lived</td>
<td>71</td>
<td>harm</td>
</tr>
<tr>
<td>22</td>
<td>plaintiff</td>
<td>47</td>
<td>lawsuits</td>
<td>72</td>
<td>inadequate</td>
</tr>
<tr>
<td>23</td>
<td>concluded</td>
<td>48</td>
<td>cfo</td>
<td>73</td>
<td>update</td>
</tr>
<tr>
<td>24</td>
<td>dismiss</td>
<td>49</td>
<td>investigation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>complaint</td>
<td>50</td>
<td>intend</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.4 BASELINE RESULTS USING SVM CLASSIFIER

In the previous section, we presented preliminary baseline results using the Naïve Bayes classifier. We also ran baseline experiments using Support Vector Machines (SVM), the main classifier used in this study. We used Waikato Environment for Knowledge Analysis (WEKA) in our experiments to train the SVM classifier to build the fraud detection model. WEKA is a machine learning toolkit that supports data mining tasks such as classification, clustering, and regression and contains visualization tools for data analysis, data preprocessing and feature selection (Witten and Frank, 2005). It is written in Java and is freely available under GNU General Public License. It uses the LIBSVM library for running support vector machines, which is written in C++ with source code also available in Java.

5.4.1 EXPERIMENTAL RESULTS AND DISCUSSION

Our baseline results based on a “bag of words” approach were much better with SVM than those achieved with the Naïve Bayes classifier. As explained earlier, in Section 5.1, the ordering of the words in a document is irrelevant for the classification task. The structure of the document is ignored under this approach; instead, the frequency of a word in a document is recorded. Each positive and negative example was represented by a document vector $x_i$, which describes the documents. Each of these documents has an associated class label $y_i$ such that $y_i \in \{-1,+1\}$ (see section 5.1 for more information on our baseline approach).

We ran our SVM experiments with the four versions of the datasets to study fraud detection as well as detection of different stages of fraud. All four versions of the datasets
are explained in detail under section 5.2.2. As discussed earlier, in this case also, we used 10-fold cross-validation to train the SVM classifier. Table 5.1 shows the split of the 10-K corpus between training and testing datasets.

It should be noted that unlike Rainbow (toolkit used to train Naïve Bayes classifier), WEKA only takes data files that are in Attribute Relation File Format (ARFF) format. Thus, 10-Ks that were downloaded from EDGAR and saved as text files could not be submitted to WEKA for processing in the raw format. Therefore, we converted these files into ARFF format before feeding them to WEKA for processing. For the baseline experiments, we created 16 ARFF files that represented the four versions of the dataset in combination with simple feature reduction techniques such as stop words removal and pruning. Figure 5.5 shows a sample of an AFRR data file where stop words list has not been applied. This file contains 261,110 features (words) and 1027 instances (405 fraudulent and 622 non-fraudulent documents). Due to the large number of features, it is not possible to show the entire ARFF file.
FIG. 5.5 - Sample of ARFF file used in WEKA

```
@relation fraud

@attribute communications numeric
@attribute corporation numeric
@attribute cable numeric
@attribute of numeric
@attribute contents numeric
@attribute part numeric
@attribute item numeric
@attribute business numeric
@attribute properties numeric
@attribute legal numeric
@attribute proceedings numeric
@attribute submission numeric

@attribute Indicator {Fraud, NoFraud}

@data
295,281,24,12641,12,157,66,880,76,88,88,4,.............................................Fraud
9,410,9,4743,8,55,36,337,21,44,18,3,.............................................Fraud
3,54,37,2457,0,22,29,101,4,9,2,1,.............................................Fraud
5,290,21,9311,6,134,65,388,46,64,36,6,.............................................Fraud
2,268,23,5528,7,84,62,149,173,69,25,6,.............................................Fraud

22,13,15,1449,1,31,63,36,4,11,13,2,......................................................NoFraud
18,43,5,831,0,18,20,31,12,11,6,......................................................NoFraud
20,7,5,773,1,22,31,36,2,3,3,2,......................................................NoFraud
1,46,80,5650,68,133,68,232,82,65,44,6,..............................................NoFraud
22,58,1,1235,0,19,25,133,2,9,6,2,......................................................NoFraud

```

112
Table 5.6 shows baseline results for the four datasets that we ran in WEKA using the SVM classifier. As explained earlier, WEKA uses the SMO algorithm to train SVMs. As discussed earlier, we used linear kernels. For performing classification in WEKA, features were normalized and were assigned weights. These results indicate that SVM performs best in majority of the datasets when we applied both pruning and the adjusted stop words list in our experiments.

**TABLE 5.6**
*Baseline Results with SVM Classifier*

<table>
<thead>
<tr>
<th>Features</th>
<th>Fraud Detection</th>
<th>Detection of Stages of Fraud</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fraud/NoFraud</td>
<td>Pre/Adv/Post</td>
</tr>
<tr>
<td></td>
<td>Version 1</td>
<td>Version 1</td>
</tr>
<tr>
<td></td>
<td>Version 2</td>
<td>Version 2</td>
</tr>
<tr>
<td><strong>Bag of Words (TF)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(w/o applying stop words and no pruning)</td>
<td>63.11 %</td>
<td>51.81 %</td>
</tr>
<tr>
<td></td>
<td>63.01 %</td>
<td>62.79 %</td>
</tr>
<tr>
<td><strong>Bag of Words (TF)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(with stop words only)</td>
<td>66.09 %</td>
<td>52.41 %</td>
</tr>
<tr>
<td></td>
<td>67.78 %</td>
<td>62.87 %</td>
</tr>
<tr>
<td><strong>Bag of Words (TF)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(with pruning only)</td>
<td>66.76 %</td>
<td>52.37 %</td>
</tr>
<tr>
<td></td>
<td>66.91 %</td>
<td>65.73 %</td>
</tr>
<tr>
<td><strong>Bag of Words (TF)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(with pruning and stop words)</td>
<td>71.67 %</td>
<td>51.83 %</td>
</tr>
<tr>
<td></td>
<td>69.11 %</td>
<td>65.81 %</td>
</tr>
</tbody>
</table>

The best classification accuracy score that we achieved with SVM is 71.67 percent, which is 14.92 percent more than the best score that we achieved with the Naïve Bayes classifier (see Table 5.3 for more information on baseline results using NB). These results support our claim that SVM is a better classifier for our domain problem and for text classification in general.
5.4.1.1 CONFUSION MATRIX RESULTS

The confusion matrix, as shown in Table 5.7, represents the typical metrics used for evaluating the performance of machine learning algorithms. This is followed by presentation and discussion of actual confusion matrix results that we obtained with baseline experiments (for the best score).

**TABLE 5.7**

*Template of a Confusion Matrix Table*

<table>
<thead>
<tr>
<th>Class</th>
<th>Predicted Positive</th>
<th>Predicted Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Positive</td>
<td>Number of True Positives</td>
<td>Number of False Negatives</td>
</tr>
<tr>
<td>Actual Negative</td>
<td>Number of False Positives</td>
<td>Number of True Negatives</td>
</tr>
</tbody>
</table>

Table 5.8 shows the confusion matrix results for the dataset with the best score of 71.67 percent. In this confusion matrix, the row is the actual class that the instance belongs to and the column is the class that the classifier predicts.

**TABLE 5.8**

*Confusion Matrix Results*

<table>
<thead>
<tr>
<th>Class</th>
<th>Fraud (Predicted)</th>
<th>NoFraud (Predicted)</th>
<th>Total (Actual)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraud (Actual)</td>
<td>168</td>
<td>237</td>
<td>405</td>
</tr>
<tr>
<td></td>
<td>True Positives</td>
<td>False Negatives</td>
<td></td>
</tr>
<tr>
<td>NoFraud (Actual)</td>
<td>54</td>
<td>568</td>
<td>622</td>
</tr>
<tr>
<td></td>
<td>False Positives</td>
<td>True Negatives</td>
<td></td>
</tr>
<tr>
<td>Total (Predicted)</td>
<td>222</td>
<td>805</td>
<td>1027</td>
</tr>
</tbody>
</table>

114
Table 5.8 shows that SVM correctly classified 736 instances out of the total of 1027 instances (71.67 percent) and incorrectly classified 291 instances (28.33 percent). The diagonal cells in the matrix show the number of 10-Ks that are correctly assigned by the classifier to their true class. For example, the first row indicates that there are 405 total instances that should be classified as belonging to the fraud class. The classifier classified 168 of them correctly and made 237 mistakes by incorrectly classifying 237 fraudulent reports as belonging to the no-fraud class. The second row indicates that there are 622 total instances that should be classified as belonging to the no-fraud class. In this case, the classifier classified 568 reports correctly and incorrectly classified 54 non-fraudulent reports as fraudulent reports. The mean absolute error amounted to 0.28.

**Discussion**

These results show a bias in the classifier towards the larger class (no-fraud). This is a common problem for data found in many real-world applications where the training dataset is dominated by instances belonging to one class while the other class (usually target class) is represented by only a few instances. The bias towards the no-fraud class is probably due to the fact that most classifier algorithms are trained to minimize the global error and hence classifiers tend to classify accurately the instances of the majority class and tend to misclassify the instances of the minority class.

Typically, the higher the class imbalance, the higher the classifier bias towards the majority class. In our case, for the best score of baseline results of 71.67 percent, the imbalance ratio, i.e., the ratio of no-fraud class instances to fraud class instances is 1.53:1, which is not very high.
When baseline results for fraud detection were compared to the random baseline, we noted that baseline experiments with SVM were able to beat the random baseline whereas Naïve Bayes was unable to beat the random baseline. For fraud detection, the random baseline would yield an accuracy of 60.56 percent simply by classifying all the documents to the largest class (no-fraud). Here, SVM was able to beat the random baseline, but not by much. One possible explanation for this is that SVM was at an early stage of learning when we did our baseline experiments but as the classifier was subsequently trained with more sophisticated features (see Chapter 6 for more information on features), we observed that, the learning accuracy of the classifier increased even for the minority class ‘fraud’. A series of experiments (see Chapter 6) showed promising results when we trained the classifier with these linguistic features.

**5.4.1.2 DETAILED ACCURACY RESULTS**

Table 5.9 shows detailed accuracy results by class. It shows the True Positive (TP) rate, False Positive (FP) rate, precision, recall and F-measure results for the dataset with the best score of 71.67 percent. For the fraud class, true positives indicate the number of fraudulent 10-Ks that are correctly classified as fraudulent, whereas false positives indicate the number of non-fraudulent 10-Ks that are incorrectly classified as fraudulent. For the no-fraud class, true positives indicate the number of non-fraudulent 10-Ks that are correctly classified as non-fraudulent and false positives indicate the number of fraudulent 10-Ks that are incorrectly classified as non-fraudulent.
TABLE 5.9
Detailed Accuracy Results

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraud</td>
<td>0.415</td>
<td>0.087</td>
<td>0.757</td>
<td>0.415</td>
<td>0.536</td>
</tr>
<tr>
<td>NoFraud</td>
<td>0.913</td>
<td>0.585</td>
<td>0.706</td>
<td>0.913</td>
<td>0.796</td>
</tr>
</tbody>
</table>

The TP rate is obtained by dividing the number of true positives by the sum of true positives and false negatives. The FP rate is obtained by dividing the number of false positives divided by the sum of true negatives and false positives.

Precision for a class is defined as the number of correct predictions of a class divided by the total number of predictions for that class, whereas recall for a class is defined as the number of correct predictions of a class divided by the total number of actual instances of that class in the dataset. The F-measure is the weighted harmonic mean of precision and recall and it measures the performance of the classifier (see Chapter 3 for more information on testing the performance of a classifier).

Discussion

These results indicate that the FP rate of 0.585 for the no-fraud class is much higher than the FP rate of 0.087 for the fraud class. The high rate of false positives for the no-fraud class is not a desirable situation as this indicates that the classifier missed 58.5 percent of the fraudulent annual reports (type I error) and misclassified them as non-fraudulent annual reports, which is more dangerous than the case where the classifier misclassifies non-fraudulent annual reports as fraudulent annual reports (type II error), thus creating a false alarm. Table 5.10 outlines the nature of Type I and Type II errors.
TABLE 5.10
Type I and Type II Errors

<table>
<thead>
<tr>
<th>Fraud Exists (Actual)</th>
<th>Fraud Predicted by the Classifier</th>
<th>Fraud Not Predicted by the Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct Classification</td>
<td>Type I Error</td>
</tr>
<tr>
<td>Fraud Doesn’t Exist</td>
<td>Type II Error</td>
<td>Correct Classification</td>
</tr>
</tbody>
</table>

These results also indicate that the TP rate of 0.415 for the fraud class is much lower than the TP rate of 0.913 of the no-fraud class. A classifier is considered superior to another if it has a higher TP rate and a lower FP rate. Here, the low TP rate of 0.415 for the fraud class is not a desirable situation, even though its FP rate is low. In general, the FP rate goes up as one attempts to increase the TP rate. This is evident in the case of the no-fraud class, which has a high TP rate of 0.913 along with high FP rate of 0.585. Here, the classifier was more liberal in the sense that it made positive predictions even when there was weak evidence, which resulted in a high TP rate but also a high FP rate.

As observed in Table 5.9, even though the recall rate for the minority class ‘fraud’ was lower (which might be due to the fact that the distribution of data is skewed), its predictive accuracy was higher. This is due to the fact that the classifier got only those instances of fraudulent annual reports correct where strong evidence was present; therefore there is a low TP rate for the fraud class, but also few false positive errors.

From our baseline experiments, we conclude that further training of the classifier is required with more sophisticated features to minimize the FP rate for both no-fraud and fraud classes as well as to maximize the TP rate for the fraud class. We discuss the next series of these experiments in detail in Chapters 6 and 7.
5.4.2 WEKA Explorer

WEKA also provides visualization tools to visualize the class distribution for each feature in the dataset. It creates a color-coded histogram for each feature and we can use these charts to examine the behavior of the extracted features. Figure 5.6 presents a sample of the attributes (features) contained in the dataset. The fraud class is represented by the color blue whereas the no-fraud class is represented by the red color. If the type of the attribute is numeric, then it also provides statistics describing the distribution of values in the dataset such as its maximum value, minimum value, mean, and standard deviation. For instance, for the numeric attribute ‘communications,’ there are 119 distinct values. Its minimum value is 0, maximum value is 541, mean is 18.479, and standard deviation is
In the case of a nominal attribute, WEKA provides each possible value that the attribute can take along with the number of instances that have that value. In our dataset, all the attributes are numeric except the attribute ‘Indicator’ which is of nominal (‘Fraud,’ ‘NoFraud’) type. In this dataset, we have 405 instances in the fraud class and 622 instances in the no-fraud class.

5.4.3 DOCUMENT-TERM MATRIX

We also analyzed the data by creating a document-term matrix. This representation also helped us construct ARFF files that were given to the SVM classifier in WEKA. The analysis of the initial document-term matrix showed a very sparse matrix. This motivated us to prune the feature space by removing all those words that occur in three or less than three documents. On further examination of the document-term matrix, we observed that the feature space consisted of many meaningless words such as ‘nacional,’ ‘tercompany,’ ‘xxxix,’ ‘jaunay,’ and ‘totion.’ An investigation of these words revealed that they have no useful meaning in the accounting domain or otherwise, and all of these words were eliminated from the feature space. These two steps helped us dramatically reduce the feature space from a total of 261,110 words to 25,270 words.

Table 5.11 illustrates a sample of one of the document-term matrices that we constructed. Each row of the matrix corresponds to the documents in the corpus and each column of the matrix corresponds to a unique term (word) that appears in the whole corpus. Each cell of the matrix indicates the term frequency TF(w, d), i.e., the number of times each word w occurs in each document d. The matrix consists of 25,270 columns (representing 25,270 unique terms) and 1,027 rows (representing a total of 405 fraudulent and 622 non-fraudulent 10-Ks). Here, frequency of each term is used as the term weight.
TABLE 5.11 – Illustration of Document-Term Matrix

<table>
<thead>
<tr>
<th>Features</th>
<th>communications</th>
<th>corporation</th>
<th>cable</th>
<th>of</th>
<th>contents</th>
<th>part</th>
<th>item</th>
<th>business</th>
<th>properties</th>
<th>legal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of Features = 25270</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADELOQ 2000</td>
<td>295</td>
<td>281</td>
<td>24</td>
<td>12641</td>
<td>12</td>
<td>157</td>
<td>66</td>
<td>880</td>
<td>76</td>
<td>88</td>
</tr>
<tr>
<td>ADPT 1998</td>
<td>9</td>
<td>410</td>
<td>9</td>
<td>4743</td>
<td>8</td>
<td>55</td>
<td>36</td>
<td>337</td>
<td>21</td>
<td>44</td>
</tr>
<tr>
<td>AIG 2002</td>
<td>3</td>
<td>54</td>
<td>37</td>
<td>2457</td>
<td>0</td>
<td>22</td>
<td>29</td>
<td>101</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>HC 2003</td>
<td>5</td>
<td>290</td>
<td>21</td>
<td>9311</td>
<td>6</td>
<td>134</td>
<td>65</td>
<td>388</td>
<td>46</td>
<td>64</td>
</tr>
<tr>
<td>HRC 2002</td>
<td>2</td>
<td>268</td>
<td>23</td>
<td>5528</td>
<td>7</td>
<td>84</td>
<td>62</td>
<td>149</td>
<td>173</td>
<td>69</td>
</tr>
<tr>
<td>XTR 1999</td>
<td>8</td>
<td>98</td>
<td>6</td>
<td>1716</td>
<td>0</td>
<td>31</td>
<td>35</td>
<td>118</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>YBTVA 2003</td>
<td>21</td>
<td>18</td>
<td>11</td>
<td>2247</td>
<td>2</td>
<td>38</td>
<td>43</td>
<td>39</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>ZNT 2001</td>
<td>2</td>
<td>20</td>
<td>22</td>
<td>1762</td>
<td>1</td>
<td>30</td>
<td>41</td>
<td>72</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>Total Number of Documents = 1027</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.5 SUMMARY

In this chapter, we presented our baseline results for the “bag of words” approach using two popular classifiers, Naïve Bayes (NB) and Support Vector Machines (SVM). When we compared our initial baseline results using NB with those obtained using SVM, we noted that SVM performed better than NB. More precisely, SVM achieved an accuracy of 71.67 percent whereas the best score achieved with NB was only 56.75 percent. We consider our baseline results with SVM very promising to support our claim that SVM is better suited to our domain problem. For both classifiers, the best overall performance in the four datasets with four different combinations was achieved with version 1 of the fraud dataset with employment of minimal feature reduction techniques such as pruning and stop words removal.
CHAPTER 6
EXPERIMENTS AND RESULTS

In this chapter we present experiments that we ran to examine the verbal content and presentation style of the annual reports using linguistic features. This chapter also describes the feature extraction process and feature selection process that we employed in tandem with the classifier results. Section 6.1 discusses features related to style (subsections 6.1.2–6.1.7) and content (subsections 6.1.8–6.1.10) as well as the feature extraction process.

Subsection 6.1.1 includes a description of the tools (DICTION 5.0 and STYLE) that we used for extracting linguistic features. Subsection 6.1.2 presents simpler linguistic features such as the average and standard deviation of the lengths of words in the document. Subsections 6.1.3 through 6.16 focus on other linguistic elements that are useful for capturing style such as voice, uncertainty markers, tone, and readability index. Subsection 6.1.7 explores linguistic elements requiring deeper analysis such as word usage, sentence formation and type-token ratio. Here we perform syntactic analysis by parsing the data to examine sentence structure. Subsections 6.1.8 through 6.1.10 examine content-related features such as keywords, bigrams, and TFIDF-based tokens.

Section 6.2 presents the feature selection process and the features that were finally selected to build the fraud model. It also provides results of the experiments that we obtained with the features and a ranking of the features by order of contribution they
make in accurately recognizing fraudulent and non-fraudulent annual reports and different stages of fraud for early detection. Finally, Section 6.3 summarizes the findings of the experiments.

6.1 STYLE AND CONTENT FEATURES

Style of a text refers to the mode of expression that the writer has employed in his or her writing to communicate the content. It basically refers to the writing characteristics such as vocabulary choices, grammatical constructions, and figures of speech that the writer uses for the purpose of expressing the material. Style denotes collective characteristics of writing that contribute to making the writing a piece of art. As Pater (1889) has observed, it is the shape that the whole takes when each and every part has fallen into its place according to the design that the writer had in mind. Pater has further observed that in the initial stages, the writer makes a lot of effort in perfecting his or her style and consciously makes lexical choices. However, once the writer has mastered a style, the style looks effortless and so spontaneous that it gives the impression that real art lies in the concealment of art. From a linguistic perspective, style implies the linguistic choices that the writer makes on a consistent basis.

We examine both the verbal content and the presentation style of the annual reports to detect fraud and stages of fraud. The features relating to content focus on the ‘what’ part of the annual report, that is, what it contains; whereas features relating to presentation style focus on the ‘how’ part of the annual report, that is, how its content is communicated.
We perform our experiments with four feature sets. The first feature set consists of eight simple surface features such as the average length of the words, the standard deviation of the word length in the document, and the average length of the sentences in the document. The second feature set consists of four features: voice, frequency of uncertainty markers, tone, and readability index. The third feature set involves deeper linguistic analysis and includes features such as word usage, formation of sentence beginnings and type-token ratio. The fourth feature set consists of content-related features such as keywords, bigrams and tf-idf words.

The pre-selection of these features was inspired by our informed reasoning and domain knowledge and rests on the speculation that the qualitative content of annual reports manifest linguistic cues that can be used for detecting fraud (see Chapter 3 for more information). As the classifier converged to higher levels of accuracy, we isolated features that had the most discriminative power in terms of detecting fraud and stages of fraud and ranked them in order of their relevance to our domain problem. Next, we provide an overview of the tools—DICTION 5.0 and STYLE—that we used for extracting linguistic features before presenting these individual features in subsections 6.1.2 through 6.1.10.

### 6.1.1 TOOLS

We primarily used two tools—DICTION 5.0 and STYLE—for extracting most of our linguistic features. In this section, we also discuss how we implemented feature processing using each of these tools.
6.1.1.1 DICTION 5.0

DICTION 5.0 is a text analysis, Windows-based commercially available software program created by Roderick Hart (2000). It uses a lexicon of 10,000 words and assigns each word to a concept classification or into a thematic category. It has five major thematic categories (Activity, Optimism, Certainty, Realism and Commonality) that are composed of combinations of 31 sub-feature categories and four calculated sub-feature categories. The scores on these sub-feature categories are aggregated to form the five major thematic categories. DICTION 5.0 can process an unlimited number of input text files and uses a series of dictionaries (built-in word-lists) to search a text. It allows for creation of custom dictionaries in addition to standard dictionaries if built-in dictionaries are not sufficient for meeting particular research needs. DICTION also outputs reports containing results about the texts it processes. For example, its processed output includes general language statistics such as total words analyzed, total characters analyzed, average word size, and number of different words. It also provides special counts of orthographic characters and high frequency words. In addition, it also reports standard dictionary totals including raw frequencies, percentages, and standardized scores for DICTION’s built-in word lists and provides totals for custom dictionaries.

Its calculated sub-feature categories include an ‘Insistence Score,’ which indicates a text’s dependence on a limited number of often-repeated words (all words occurring three or more times that function as nouns or noun-derived adjectives), an ‘Embellishment Score’ (the ratio of descriptive to functional words), a ‘Variety Score’ (a measure of linguistic dispersion), and a ‘Complexity Score’ (a measure of word size). Standardizing all previous dictionary scores it composes five overall measures of semantic features.
which, taken together, provide the most general understanding of the given text (Hart, 2000).

The processing of features using DICTION 5.0 involved the following steps:

1. We created two master project files in DICTION—‘Fraud’ and ‘No-Fraud.’

2. We imported the fraudulent and non-fraudulent annual reports into DICTION under the two projects—‘Fraud’ and ‘No-Fraud’—respectively. For this purpose, all fraudulent and non-fraudulent annual reports were converted into standard text files (with .txt extension) as DICTION only accepts text files.

3. We created custom dictionaries for extracting some of our features—uncertainty markers and tone. Since DICTION does not allow access to its proprietary built-in dictionaries, it was necessary to create custom dictionaries to ensure we know what is being coded. Since its built-in dictionaries are more geared towards social sciences, this ensured that the accounting domain specific vocabulary was not ignored in the analysis. For creating custom dictionaries we identified words that were consistent with the conceptual definition of each of these features as defined in this study. Once this process of operationalization of features was complete, we stored the custom dictionaries in program-defined dictionaries. We then activated these custom dictionaries such that they were invoked each time DICTION 5.0 software analysis was run on the textual content of annual reports.

4. Once all the fraudulent and non-fraudulent annual reports were loaded into ‘Fraud’ and ‘No-Fraud’ projects, respectively, all the reports in each project were analyzed without requiring combining all the documents into a single file (for example, for the first version
of the fraud dataset there were 405 fraudulent text files and 622 non-fraudulent text files in total).

5. We then repeated this process for other datasets and ran DICTION software on different corpora (two versions of fraud detection dataset and one version of detection of stages of fraud dataset).

6. DICTION’s output reports the results of file processing and is divided into three views—File View, Data View, Numeric View. The results for each annual report were analyzed by clicking on its name in the File View pane. The results of custom dictionaries were reported right after the standard dictionary scores in DICTION output.

7. A vector of category counts—positive words count, negative words count, and hedge words count was derived from DICTION for each annual report in the corpora. We then normalized these individual category counts for the length of the annual reports before running classifier experiments with them.

6.1.1.2 STYLE

STYLE is a UNIX command-line based GNU program that analyzes the surface characteristics of the writing style of a document, including sentence length and type, word usage and other readability measures and provides a stylistic profile of writing at the word and sentence level (Cherry and Vesterman, 1991). It reports the number of characters, sentences, and paragraphs; the average word, sentence, and paragraph length; the number of short sentences and long sentences; and voice usage. STYLE defines a sentence as a sequence of words ending in one of: period, double colon, question mark or exclamation mark. A paragraph consists of two or more new line characters.
It also provides readability scores based on a variety of readability tests such as Kincaid Formula, Automated Readability Index, Coleman-Liau Formula, Flesch Reading Ease, Robert Gunning’s Fog Index, Lix Formula, and SMOG-Grading. The readability indices reported by STYLE are based on measures of sentence and word lengths.

In addition, STYLE provides word usage summary and tells the number of verbs, with a breakdown by type, and a breakdown on types of sentence beginnings. The word usage counts are intended to help identify excessive use of particular parts of speech. STYLE reports on counts of coordinating as well as subordinating conjunctions. Coordinating conjunctions join grammatically equal sentence fragments, such as a noun with a noun, a phrase with a phrase, or a clause with a clause. Coordinating conjunctions are "and," "but," "or," "yet," and "nor." Subordinating conjunctions connect clauses of unequal status. A subordinating conjunction links a subordinate clause, which is unable to stand alone, to an independent clause. Examples of subordinating conjunctions are "because," "although," and "even if." STYLE also outputs sentences containing nominalizations, or verbs that that have been transformed into nouns, such as “deduction” (from the verb “deduct”), “judgment” (from the verb “judge”).

The processing of features using STYLE involved the following steps:

1. Since STYLE only reads plain text input either from a file or the standard input, we converted all the documents in the two corpora (fraud and no-fraud) to text documents.
2. It should be noted that in order to accurately capture style-related features, no preprocessing such as conversion of upper to lower case, removal of punctuation, etc. was performed on the dataset.

3. Using a shell script, STYLE was applied to each text file in the two corpora (fraud, no-fraud) and output was accumulated in a log file for further processing.

4. Log files were then imported onto an excel spreadsheet for further analysis. Another shell script was written to automate this step.

5. This entire process was repeated for all the datasets (see Chapter 4 for more information on datasets used in this study for fraud detection and detection of stages of fraud).

### 6.1.2 SIMPLE SURFACE FEATURES

In order to examine the presentation style of the annual reports, we first looked at eight surface-oriented features. Information on these features was collected from the text itself and did not require grammatical annotation of the text. Here, we did not care about the logical structure of a particular sentence; rather, we were interested in the average, standard deviation, and relative percentages of these features. Even though these features are often called simple surface features, their importance is undeniable. Prior research in the area of stylometry (Forsyth and Holmes, 1996) has reported good results with several of these features.

It should be noted that we did not include raw frequencies of basic surface features such as the number of words, the number of sentences, and the number of paragraphs in this feature set. This is due to the fact that annual reports varied in length from one company
to another company and from one year to another year for the same company in the two corpora. Using raw count of number of words in an annual report would not provide a good representation of the data as longer annual reports typically tend to contain a higher count of words than the shorter annual reports. Thus, in this study we used either percentages of features or features that were normalized by the length of the document. The eight simple surface features used are as follows:

1. Average length of the words (in terms of characters) in the document
2. Standard deviation of word lengths in the document
3. Average length of the sentences (in terms of words) in the document
4. Standard deviation of sentence lengths in the document
5. Percentage of short sentences (consisting of at most 30 words)
6. Percentage of long sentences (consisting of at least 60 words)
7. Average length of the paragraphs (in terms of sentences) in the document
8. Standard deviation of paragraph lengths in the document

We believe that in cases of fraud, annual reports often contain informative descriptive statements obscured by a pile of uninformative verbiage; non-fraudulent annual reports, on the other hand, are usually expressed clearly and economically. The presumption is that features such as average length and standard deviation of words and sentences can provide linguistic cues for detecting fraud.

Table 6.1 presents lexical statistics on the surface features obtained by STYLE for a sample fraudulent and non-fraudulent annual report. We gathered similar statistics on
these surface features for all the 10-Ks in the two corpora (fraud, no-fraud) for the three datasets (see Chapter 4 for more information on the datasets).

**TABLE 6.1**

*Simple Surface Features for a Sample Fraudulent and Non-Fraudulent Annual Report*

<table>
<thead>
<tr>
<th>Surface Features</th>
<th>Fraudulent 10-K</th>
<th>Non-Fraudulent 10-K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Length of the Words (in terms of characters)</td>
<td>5.36</td>
<td>4.93</td>
</tr>
<tr>
<td>Standard Deviation of Word Lengths</td>
<td>1.7</td>
<td>1.5</td>
</tr>
<tr>
<td>Average Length of the Sentences (in terms of words)</td>
<td>31.1</td>
<td>17.5</td>
</tr>
<tr>
<td>Standard Deviation of Sentence Lengths</td>
<td>11.2</td>
<td>5.3</td>
</tr>
<tr>
<td>Percentage of Short Sentences (at most 30 words)</td>
<td>60%</td>
<td>55%</td>
</tr>
<tr>
<td>Percentage of Long Sentences (at least 60 words)</td>
<td>22%</td>
<td>14%</td>
</tr>
<tr>
<td>Average Length of the Paragraphs (in terms of sentences)</td>
<td>3.7</td>
<td>2.3</td>
</tr>
<tr>
<td>Standard Deviation of Paragraph Lengths</td>
<td>1.9</td>
<td>1.4</td>
</tr>
</tbody>
</table>

### 6.1.3 VOICE

As discussed earlier, for analyzing the style of the annual reports, another particularly relevant feature that we wanted to examine was use of voice. The voice is that form of a verb which shows whether what is denoted by the subject of the sentence does something (active) or has something done to it (passive). There are two voices in English grammar: active voice and passive voice. Consider two simple sentences listed below:

(1) Jane prepared the cake.

(2) The cake was prepared by Jane.
When we analyze these two sentences using transformational rules of grammar, sentence (1) can be represented as:

Jane prepared the cake

where NP = Noun Phrase

VP = Verb Phrase

Structure Description (SD) of sentence (1) can be defined as:

\[ SD = 1 + 2 + 3 \]

After transformation, sentence (1) becomes:

The cake was prepared by Jane

This change is called Structural Change (SC) and can be written as

\[ SC = 3 + was + 2 + by + 1 \]

In other words, a transformation rule has moved the constituents and inserted ‘was’ and ‘by’ in sentence (2). In practice, the passive transformation is more complex but it has been kept simple for illustrating this rule. Consider the following sentences selected from annual reports (initial structure of the sentence is in bold to illustrate active and passive voice).
1. The company has made a substantial commitment to the technological development of the Systems and is aggressively investing in the upgrade of the technical capabilities of its cable plant in a cost efficient manner.

2. The proceeds from these issuances were used to repay short-term borrowings and fund the cash needs of the U.S. operations.

3. These facilities were established in September 2001 and August 2003, respectively, with a syndicate of lenders.

Sentence (1) gives an example of an active voice sentence, where the person or thing represented by the subject performs the action represented by the verb. Sentences (2) and (3) illustrate examples of passive voice sentences in which the object occupies the subject position of the sentence, and the subject is left out altogether (truncated passive). In the case of full passive, the subject is placed in an agent phrase at the end of the sentence. The use of the truncated passive is more common than use of the full passive (Williams, 1994). In these examples, the main verb appears in the past participle form and comes after some form of the verb “to be” (see Sentences 2 and 3).

It can be observed from these illustrations that the passive voice typically uses more “weak” words (including abstract words such as “is,” “am,” “are,” “was,” “were,” “being” and “been”; demonstrative pronouns such as “the”; and prepositions such as “by” and “of”) in comparison to the active voice, which uses “strong” words (including concrete nouns, powerful verbs, and vivid adjectives). For instance, “to-be” verbs and prepositions such as “by” and “of” do not add much to the life of the sentence.
The passive voice is typically used when the speaker wants to hide the agent or obscure what occurs and as a result the sentence becomes unclear. In this type of situation, the passive voice is used in a deceptive way to avoid disclosing uncomfortable news. Linguistic studies have shown that people are better able to remember material they read in the active voice than the same material in the passive voice. They noted that in the case of the active voice, the mind remains geared toward a “Subject-Verb-Object” pattern, whereas passive voice sentences derail that mental process of retention (Coleman, 1965).

It should be further noted that the use of the passive voice often requires up to 30% to 40% more words to express the same idea as compared to the active voice. Active voice sentences tend to be more concise than passive voice sentences.

We believe that in cases of fraud, management deliberately tries to shift responsibility away from itself and makes more use of passive voice sentences than active voice sentences. In contrast, management makes more use of active voice sentences when it wants to take credit for the positive outcomes by attributing such outcomes to its own actions and wants to accept accountability for its actions. We used GNU software STYLE to extract percentages of passive voice and active voice sentences for all documents in the three datasets (see Chapter 4 for more information on the datasets).

### 6.1.4 UNCERTAINTY MARKERS

The uncertainty markers, also called modal verbs or hedge words, include such words as “can,” “could,” “may,” “might,” “shall,” “should,” “will,” “would,” “must,” “ought,” “probably,” “likely,” “rather,” “sort of,” “kind of” and so on. Consider the following
sentences that have been extracted from the annual reports to demonstrate use of these words (uncertainty markers are indicated by bold underlined words).

1. In addition, various legislative and regulatory proposals under consideration from time to time by Congress and various federal agencies may materially affect the Company’s existing and anticipated businesses.

2. These decisions have been somewhat inconsistent and, until the U.S. Supreme Court rules definitively on the scope of cable television’s First Amendment protections, the legality of the franchising process of various specific franchise requirements is likely to be in a state of flux.

3. The Company believes that the provision of video programming by telephone companies in competition with the Company’s existing operations could have an adverse effect on the Company’s financial condition and results of operations. At this time, the impact of any such effect is not known or estimable.

Uncertainty markers have been extensively used in the literature to study style, expression, affect, and attitude in text (Lackoff, 1973; Glover and Hirst, 1996; Uzuner and Katz, 2005b; Rubin et al., 2006). In linguistics, uncertainty markers are often associated with weakening and deintensification of the message. Several studies relating to deception analysis have also used uncertainty markers to isolate cases of deception. These words have been shown to function as a subtle means to avoid responsibility and evade the truth.
We believe that in the case of fraud, management purposely employs more uncertainty markers to make annual reports vague and ambiguous. Thus, instead of providing correct and accurate information, management often obscures facts with intent to hide the true picture. We expect that the examination of uncertainty markers in the annual reports would help to unravel some of these linguistic predictors of fraud.

We used Windows-based DICTION 5.0 commercially available software to extract the frequency of uncertainty markers in our datasets. This software allows the creation of custom dictionaries, which can be used to search text documents. Since there was no built-in dictionary in DICTION to provide a count of uncertainty markers, we created a custom dictionary called ‘HEDGEWORDS.DIC,’ and ran all the files in each of the three datasets to compute the frequency of uncertainty markers in them (see Chapter 4 for more information on datasets). We then scaled these raw frequency scores of occurrences of uncertainty markers by the total number of words in an annual report to compute normalized scores for all the 10-Ks in the three datasets.

**6.1.5 READABILITY INDEX**

A readability index is a measure of the ease or difficulty of reading and understanding a piece of text. It takes into consideration surface characteristics of the text such as the number of words in the sentences, the number of characters per word or the number of syllables per word to measure readability and comprehensiveness of the text. There are different measures available to compute readability grades such as ‘Flesch-Kincaid Grade Level,’ ‘Automated Readability Index,’ ‘Coleman-Liau Index,’ ‘Flesch Reading Ease Score,’ ‘Gunning Fog Index,’ ‘Lix Formula,’ and ‘SMOG-Grading.’ We compared the
scores on these readability grades for fraudulent 10-Ks and non-fraudulent 10-Ks to
detect fraud. Furthermore, we compared the readability grade scores of 10-Ks for pre-
fraud, adv-fraud, and post-fraud periods to detect stages of fraud. We briefly describe
these readability grades in Table 6.2.
TABLE 6.2
Definitions of Readability Grades

<table>
<thead>
<tr>
<th>Readability Grade</th>
<th>Formula/Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(syll = syllables, wds=words, chars=characters, sent=sentences)</td>
</tr>
<tr>
<td>Flesch-Kincaid Grade Level</td>
<td>Kincaid = 11.8*(syll/wds)+0.39*(wds/sent)–15.59</td>
</tr>
<tr>
<td></td>
<td>The Kincaid formula translates the 0-100 score to a U.S. grade level and indicates the number of years of education generally required to understand the text. For instance, a grade level score of 0 is considered very easy and a score of 12 is deemed to be very hard.</td>
</tr>
<tr>
<td>Automated Readability Index (ARI)</td>
<td>ARI = 4.71*(chars/wds)+0.5*(wds/sentences)–21.43</td>
</tr>
<tr>
<td></td>
<td>The ARI score indicates the U.S. grade level needed to comprehend the text. ARI is typically higher than the Flesch-Kincaid Grade Level and Coleman-Liau Index but lower than the Flesch Reading Ease Score.</td>
</tr>
<tr>
<td>Coleman-Liau Index</td>
<td>Coleman-Liau = 5.89*(chars/wds)–0.3*(sent/(100*wds))–15.8</td>
</tr>
<tr>
<td></td>
<td>Coleman-Liau output approximates the U.S. grade level. It considers characters instead of syllables per word and usually gives a lower score than the Flesch-Kincaid Grade Level, ARI and Flesch Reading Ease Score for technical documents.</td>
</tr>
<tr>
<td>Flesch Reading Ease Score</td>
<td>Flesch Index = 206.835-1.015*(wds/sent)-84.6*(syll/wds)</td>
</tr>
<tr>
<td></td>
<td>The Flesch index is usually between 0 (hard) and 100 (easy) with higher scores indicating the text is easier to read whereas lower scores indicate a greater reading difficulty. Typically, scores from 0-30, 30-50, and 50-60 are interpreted as very difficult, difficult and fairly difficult to read, respectively.</td>
</tr>
<tr>
<td>Gunning Fog Index</td>
<td>Fog Index = 0.4*(wds/sent+100*((wds&gt;=3syll)/wds))</td>
</tr>
<tr>
<td></td>
<td>The Fog index value denotes a school grade. A value above 12 indicates that the text is too hard for average people to read.</td>
</tr>
<tr>
<td>Lix Formula</td>
<td>Lix = (wds/sent)+(100*(wds&gt;=6char))/wds</td>
</tr>
<tr>
<td></td>
<td>The Lix formula takes into consideration only long words, i.e., words containing six or more characters. Typically, scores from 0-24, 25-34, 35-44, 45-54, 55 and above are interpreted as very easy, easy, standard, difficult, and very difficult to read, respectively.</td>
</tr>
<tr>
<td>SMOG-Grading</td>
<td>SMOG-Grading = square root of (((wds&gt;=3syll)/sent)*30)+3</td>
</tr>
<tr>
<td></td>
<td>The SMOG-Grading indicates a school grade, which is a rough measure of how many years of schooling it would take someone to understand the content.</td>
</tr>
</tbody>
</table>
Several studies (Courtis, 1986; Baker and Kare, 1992; Smith and Taffler, 1992; Subramanian et al., 1993) have used readability tests to examine the relationship between the readability of annual reports and corporate failures, or corporate profitability. However, these studies show mixed results (see Chapter 3 for more information on these studies). Many studies in computational linguistics have also used readability indices to examine readability of texts (see, for example, Mikk, 1995; Das and Roychoudhury, 2006).

We believe that in cases of fraud, management intentionally makes annual reports harder to read by making use of longer and uncommon words and more complex syntactic sentence structures. Furthermore, it is particularly important to do readability analysis in the case of a qualitative examination of annual reports as these reports are read by a wide number of users and are the most important vehicle through which companies communicate their past performance and future prospects. We examine readability indices under the assumption that fraudulent annual reports are harder to read and comprehend in comparison to non-fraudulent annual reports.

We used GNU software STYLE to compute several readability indices for the annual reports in the corpus. Table 6.3 shows readability scores using different measures for a sample fraudulent and non-fraudulent annual report. We collected information on all of these readability measures to analyze the corpus in terms of its level of readability and come to a more definite conclusion as to which measure is a better predictor for our domain problem. We finally chose ‘Flesch-Kincaid Grade Level’ and ‘Flesch Reading Ease Score’ to build the fraud model.
TABLE 6.3

Readability Grades for a Sample Fraudulent and Non-Fraudulent Annual Report

<table>
<thead>
<tr>
<th>Readability Grades</th>
<th>Fraudulent 10-K</th>
<th>Non-Fraudulent 10-K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flesch-Kincaid Grade Level</td>
<td>16.8</td>
<td>12.5</td>
</tr>
<tr>
<td>Automated Readability Index</td>
<td>19.6</td>
<td>14.7</td>
</tr>
<tr>
<td>Coleman-Liau Index</td>
<td>18.2</td>
<td>13.6</td>
</tr>
<tr>
<td>Flesch Reading Ease Score</td>
<td>35.6</td>
<td>45.8</td>
</tr>
<tr>
<td>Gunning Fog Index</td>
<td>20.3</td>
<td>16.3</td>
</tr>
<tr>
<td>Lix Formula</td>
<td>64.7</td>
<td>56.4</td>
</tr>
<tr>
<td>SMOG-Grading</td>
<td>16.3</td>
<td>14.1</td>
</tr>
</tbody>
</table>

6.1.6 TONE

The tone defines the semantic orientation of a text. Simply put, it refers to the emotional attitude expressed in a text and can be measured by examining the lexical choices made by the writer, i.e., words chosen to indicate polarity of the tone. It can be classified as positive, negative, objective, subjective, pessimistic, or optimistic. Even though there is a growing interest in identifying measures for ascertaining tone, polarity, and sentiment of a text, it is still challenging to determine the overall tone of a text, as both negative and positive aspects of the tone may be present in it.

We classify tone as positive (optimistic) and negative (pessimistic) and measure it in terms of frequency count of positive and negative words scaled by the total number of words in an annual report. For this, we develop two categories of tone--positive and negative--based on the prior work of researchers in this context (Abrahamson and Amir, 1996; Smith and Taffler, 2000; Hand et al., 2001; Henry, 2005). Abrahamson and Amir (1996) examined the tone of the chairman’s letter to investors in annual reports. Smith and Taffler (2000) determined the relationship between negative words in the CEO letter
to shareholders and company failure. Henry (2005) studied the relationship between tone and investor reaction to earnings announcement and concluded that the predictive accuracy of the model improved by including tone as one of the features. Table 6.4 presents positive and negative tone category words. The original list of positive and negative words was adjusted with words found in the fraud corpus.

**TABLE 6.4**

*Tone Category Words*

<table>
<thead>
<tr>
<th>Positive (Optimistic) Words</th>
<th>Negative (Pessimistic) Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Above</td>
<td>Adverse</td>
</tr>
<tr>
<td>Absolutely</td>
<td>Apparently</td>
</tr>
<tr>
<td>Accomplish</td>
<td>Bad</td>
</tr>
<tr>
<td>Achieve</td>
<td>Barely</td>
</tr>
<tr>
<td>Astronomical</td>
<td>Below</td>
</tr>
<tr>
<td>Beat</td>
<td>Challenge</td>
</tr>
<tr>
<td>Better</td>
<td>Decline</td>
</tr>
<tr>
<td>Certain</td>
<td>Decrease</td>
</tr>
<tr>
<td>Completely</td>
<td>Delinquent</td>
</tr>
<tr>
<td>Definite</td>
<td>Depressed</td>
</tr>
<tr>
<td>Delighted</td>
<td>Deteriorate</td>
</tr>
<tr>
<td>Deliver</td>
<td>Difficult</td>
</tr>
<tr>
<td>Double</td>
<td>Disappoint</td>
</tr>
<tr>
<td>Encouraged</td>
<td>Down</td>
</tr>
<tr>
<td>Enhance</td>
<td>Downturn</td>
</tr>
<tr>
<td>Enjoy</td>
<td>Drop</td>
</tr>
<tr>
<td>Enormous</td>
<td>Fail</td>
</tr>
<tr>
<td>Entire</td>
<td>Fall</td>
</tr>
<tr>
<td>Exceed</td>
<td>Fluctuation</td>
</tr>
<tr>
<td>Excellent</td>
<td>Hardly</td>
</tr>
<tr>
<td>Extreme</td>
<td>Hurdle</td>
</tr>
<tr>
<td>Fortunate</td>
<td>Indefinite</td>
</tr>
<tr>
<td>Fully</td>
<td>Inferior</td>
</tr>
<tr>
<td>Good</td>
<td>Investigation</td>
</tr>
<tr>
<td>Greater</td>
<td>Lawsuit</td>
</tr>
<tr>
<td>Grow</td>
<td>Litigation</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Improve</td>
<td>Nasty</td>
</tr>
<tr>
<td>Immense</td>
<td>Negative</td>
</tr>
<tr>
<td>Increase</td>
<td>Obstacle</td>
</tr>
<tr>
<td>Larger</td>
<td>Penalty</td>
</tr>
<tr>
<td>Leader</td>
<td>Poor</td>
</tr>
<tr>
<td>Leading</td>
<td>Possibly</td>
</tr>
<tr>
<td>More</td>
<td>Risk</td>
</tr>
<tr>
<td>Nice</td>
<td>Roughly</td>
</tr>
<tr>
<td>Opportunity</td>
<td>Scarcely</td>
</tr>
<tr>
<td>Perfect</td>
<td>Shortfall</td>
</tr>
</tbody>
</table>
We believe that in cases of fraud, the tone of the annual report tends to be more negative as management consciously chooses selective accounting language to conceal fraud and uses more words with negative financial connotations. Conversely, tone is positive with extensive use of optimistic vocabulary for a non-fraudulent company.

We used Windows-based DICTION 5.0 commercially available software to extract frequency counts of negative and positive words in our datasets. Since, there was no built-in dictionary in DICTION to provide a count of tone category words specific to fraud, we created two custom dictionaries called ‘NEGATIVE.DIC’ and ‘POSITIVE.DIC,’ containing negative category words and positive category words, respectively. We ran all of the annual reports through DICTION to get a raw count of negative and positive category words in them. We then scaled these raw frequency scores of positive and negative tone by the total number of words in an annual report to compute normalized tone scores for all the 10-Ks in the three datasets.
6.1.7 DEEPER SURFACE FEATURES

We believe that companies that commit fraud will have high motivation to lie successfully in order to avoid exposing themselves. This in turn may influence the choice of language and style of writing employed by the management of these companies. In order to learn about management’s underlying motives, we wanted to investigate lexical choices in depth. Prior research suggests that markers of linguistic style—articles, pronouns, prepositions, and conjunctions—are, in many respects, as meaningful as specific nouns and verbs in telling what people are thinking and feeling (Dulaney, 1982; Colwell et al., 2002; Zhou et al., 2002). In deeper surface features, we examined the vocabulary frequencies (of proper nouns, pronouns, conjunctions, prepositions, nominalizations, verb types, sentence openers) in addition to the vocabulary richness (type-token ratio) in order to explore the underlying grammatical relations and identify patterns of usage in writings of the two corpora. Zhou et al. (2002) found that high variety index (type-token ratio) is associated with deception. They noted that in cases of deception, the writer uses superfluous and meaningless language to give the impression of completeness.

The motivation to examine proper nouns and pronouns in this feature set stemmed from analysis of the top discriminative words (see Chapter 5 for a sample of top discriminative words by information gain). The top discriminative words consisted of proper nouns (e.g., Argentina, Manhattan, Brazil, Venezuela, Philippines), and pronouns (e.g., my, us), which led us to exploit syntactical variation in the two corpora. Pronouns add cohesiveness and connectivity to a document by providing contextual references to nouns and noun phrases. Even though there are other mechanisms for such connections,
documents with fewer pronouns generally lack cohesiveness and fluidity and tend to be wordy and disconnected. Nominalizations have been associated with complicated clause construction and less use of active verbs, thus reducing the effectiveness of the sentence in several ways. Use of too much nominalization in a document can be abstract and difficult to understand. Another agreed upon principle of style is variety in sentence openers. We examined the sentence openers by looking at the part of speech of the first word in the sentence.

Simple surface features examined earlier indicated that there were structural differences between fraudulent and non-fraudulent annual reports. However, the features examined in the simple surface feature set are under the conscious control of the writers. On the contrary, Yule (1938) found that some of the useful features that represent the specific style are those that the writer does unconsciously. Holmes (1994) noted that features such as the use and frequency of function words (determiners, conjunctions and prepositions etc.) were useful for characterizing the style, as they were not under the conscious control of the writer.

In deeper surface features, as discussed earlier, we examined a variety of syntactical features such as distribution of the parts of speech types and verb types to explore the constructional features of writing style that writer does unconsciously. Here, we were interested in examining how different lexical categories function to create the desired communicative effect. The identification of the grammatical function of a word in a sentence will indicate variation in lexical choice and will reflect interesting variation in overall style. Furthermore, differences between the vocabularies of the two sets of corpora (fraud, no-fraud) might also reveal important clues about the underlying motives.
of the management. At this stage, we wanted to see if deeper surface features would contribute significantly to the classification accuracy. Furthermore, we wanted to test if syntactical differences between fraudulent and non-fraudulent reports were large enough to motivate future work in this area.

As noted earlier, the examination of the syntactic categories involved deeper linguistic analysis. For instance, the documents were searched for the presence of “to be” verbs and auxiliary forms of verbs; for use of nominalizations; for conjunction use, preposition use, and proper noun use; and for sentences beginning with interrogative pronouns. We also computed type-token ratios for each of the documents in the two corpora. In total, we looked at 14 features in this set. We believed that investigation of these features would help us in distinguishing simple styles of annual reports from the ponderous ones. For instance, two sentences that have similar surface characteristics may have the same or different underlying structure. Consider the two sentences listed below:

(1) The forecast for today is bright and sunny.

(2) Tom feels bright and fresh in the morning.

The quantitative measurement (e.g., total number of words, total number of characters, average word length) of the surface features for the two sentences (1) and (2) is similar. However, if we look at the phrase structure of these two sentences, it is noted that these two sentences are quite different.

This feature set consists of the following features:

1. “To be” verbs – These verbs act as linking verbs and are used to join a subject with a noun or adjective complement, to precede the ing-form of an action verb,
and to precede the past participle and present participle of a transitive verb (e.g., “is,” “am,” “are,” “was,” “have been,” “being,” “to be”).

2. Auxiliary verbs – Auxiliary verbs precede the main verb (e.g., “to have,” “to be,” “to do”) and modify the mood of a verb.

3. Conjunctions – Conjunctions connect words or parts of sentences (e.g., “and,” “or,” “but,” “not,” “yet,” “so,” “for”).

4. Proper Nouns – Proper nouns are those words that refer uniquely to a specific person, place, or thing such as “Iowa” or “Alaska.”

5. Pronouns – Pronouns are a subgroup of nouns and act as substitutes for nouns (e.g., “he,” “they,” “we”).

6. Prepositions – Prepositions express relationships between other words, usually nouns, including a relationship of time or space (e.g., “with,” “on,” “in,” “to,” “of,” “beyond,” “over,” “under,” “until,” “before,” “through”).

7. Nominalizations - Nominalization refers to the use of a verb or an adjective as a noun, with or without morphological transformation (e.g., “change,” “good,” “bad”).

8. Sentence beginnings with a pronoun

9. Sentence beginnings with an interrogative pronoun – An interrogative pronoun is a pronoun used in order to ask a question (e.g., “what,” “which,” “who”).

10. Sentence beginnings with an article – Articles precede the noun they modify (e.g., “a,” “an,” “the”).
11. Sentence beginnings with a subordinating conjunction – Subordinating conjunctions modify a dependent or subordinate clause in some way and join it with an independent clause. Sometimes, a subordinate conjunction is a phrase (“as if,” “as soon as,” “in order to,” “even though”) rather than a single word (“if,” “when,” “whenever,” “although,” “while,” “because”).

12. Sentence beginnings with a conjunction

13. Sentence beginnings with a preposition

14. Type-token ratio – This ratio is a measure of the number of different word types and is computed by dividing the number of different word types by the number of word tokens in a document.

We used GNU software STYLE and Windows-based DICTION 5.0 commercially available software to extract frequencies of the deeper surface features. We then computed scaled frequencies of these features by normalizing them for the length of an annual report. Features relating to word usage and sentence structures were scaled by the total number of words contained in an annual report and the total number of sentences contained in an annual report, respectively.

In order to compare the lexical richness of fraudulent and non-fraudulent annual reports, we examined the type-token ratio. The type-token ratio determines the relationship between the total number of different words and the total number of words contained in a document. Since, type-token ratio can be misleading for documents of varying lengths, we normalized the type-token ratio by dividing the text of each annual report into same size chunks and then computed the average of type-token ratios, calculated across the
entire annual report. We used DICTION 5.0 to compute type/token ratio. Table 6.5 presents normalized scores of these features for a sample fraudulent and non-fraudulent annual report.

**TABLE 6.5**

*Normalized Scores of Deeper Surface Features for a Sample Fraudulent and Non-Fraudulent Annual Report*

<table>
<thead>
<tr>
<th>Deeper Surface Features</th>
<th>Fraudulent 10-K</th>
<th>Non-Fraudulent 10-K</th>
</tr>
</thead>
<tbody>
<tr>
<td>“To be” verbs</td>
<td>12.47%</td>
<td>2.31%</td>
</tr>
<tr>
<td>Auxiliary verbs</td>
<td>6.78%</td>
<td>0.61%</td>
</tr>
<tr>
<td>Conjunctions</td>
<td>5.87%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Proper Nouns</td>
<td>21.99%</td>
<td>5.76%</td>
</tr>
<tr>
<td>Pronouns</td>
<td>1.79%</td>
<td>1.81%</td>
</tr>
<tr>
<td>Prepositions</td>
<td>12.64%</td>
<td>11.21%</td>
</tr>
<tr>
<td>Nominalizations</td>
<td>5.87%</td>
<td>3.89%</td>
</tr>
<tr>
<td>Sentence beginnings with pronoun</td>
<td>4.59%</td>
<td>3.28%</td>
</tr>
<tr>
<td>Sentence beginnings with interrogative pronoun</td>
<td>0.32%</td>
<td>0.08%</td>
</tr>
<tr>
<td>Sentence beginnings with article</td>
<td>13.85%</td>
<td>22.38%</td>
</tr>
<tr>
<td>Sentence beginnings with subordinating conjunction</td>
<td>4.93%</td>
<td>3.12%</td>
</tr>
<tr>
<td>Sentence beginnings with conjunction</td>
<td>0.32%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Sentence beginnings with preposition</td>
<td>9.60%</td>
<td>12.96%</td>
</tr>
<tr>
<td>Type-token ratio</td>
<td>0.619</td>
<td>0.324</td>
</tr>
</tbody>
</table>

**6.1.8 KEYWORDS**

In our keyword-based approach to classification, we selected the top 100 words with the highest information gain in the training corpus as fraud indicators (see Chapter 5 for a sample of such words). This list of indicators is composed of discriminative words that have the highest predictive capability for detecting fraud on the basis of information gain scores. Similarly, we used a list of the top 100 discriminative words by information gain for detecting stages (‘pre-fraud,’ ‘adv-fraud’) of fraud for early detection.
We used Windows-based DICTION 5.0 commercially available software to extract frequencies of these keywords in our datasets. This software allows the creation of custom dictionaries, which can be used to search text documents. In DICTION, we created two custom dictionaries called ‘FRAUD_KEYWORDS.DIC’ and ‘FRAUD_STAGES_KEYWORDS.DIC,’ containing the top 100 discriminative words by information gain for detecting fraud and the top 100 discriminative words by information gain for predicting levels of fraud, respectively. We ran all the files in each of the three datasets to compute the frequency of keywords in them (see Chapter 4 for more information on the datasets). We then normalized these raw frequency counts of occurrence of keywords by the total number of words in an annual report. It should be noted that we did not include the most frequent words in our keywords list, as frequent words typically tend to be function words.

6.1.9 TFIDF WEIGHTED TOKENS

We used the Weighted-Term Frequency Inverse Document Frequency (TFIDF) measure to evaluate how central a content word is in the fraud and no-fraud corpus. Under TFIDF weighting, each term in each document is assigned a composite weight. TFIDF is a popular statistical measure, which assigns the highest weight to a term when it occurs many times within a small number of documents and the lowest weight when the term occurs in almost all documents. As a consequence, features that have a high predictive power get a high weight regardless of their absolute frequencies.

Under this measure, words with high TFIDF values are considered important in a document and words that get a very small weight are considered useless for the
categorization problem. TFIDF values are computed by multiplying Term Frequency (TF) with Inverse Document Frequency (IDF). TF measures the importance of a term by how frequently it appears in a document and is computed by dividing the number of times a given term appears in a document by the total number of terms in the document. On the other hand, IDF is a general measure of the discriminative power of the term over a collection of documents and measures how infrequent a term is in the corpus. IDF is computed by taking the log of the total number of documents divided by the total number of documents containing the term. Thus, the IDF of a rare term tends to be high, whereas the IDF of a frequent term tends to be low. IDF is known to improve the precision, whereas TF is known to improve the recall. TFIDF is given by the formula:

$$\text{TFIDF} = \text{TF} \times \text{IDF}$$

$$= \frac{\text{number of times a given term appears in a document}}{\text{total number of terms in a document}} \times \log\left(\frac{\text{total number of documents in the corpus}}{\text{total number of documents containing the term}}\right)$$

Here, we constructed a TFIDF weighted document-term matrix. In this matrix, we used TFIDF weighting to normalize the raw scores of the document-term matrix presented in Chapter 5. These normalized scores vary in the range (0,1). A high score of TFIDF is achieved when the score of TF is high and the document frequency score of the term in the whole corpus is low. Each row of the matrix corresponded to the documents in the two corpora (fraud, no-fraud). Each column of the matrix corresponded to the unique
term that appeared in the whole corpus. The last column of the matrix indicated whether a particular document belongs to the fraud class or to the no-fraud class.

6.1.10 BIGRAMS

Bigrams are a special case of n-grams and can consist of a sequence of two characters, two words, or two syllables. Thus, in the case of bigrams, the feature vector consists of pairs of words instead of single words such as ‘sarbanes oxley’ and ‘generally accepted.’ We extracted bigrams collocating with the keywords (see Chapter 5 for a sample of keywords). We extracted bigrams before performing any kind of preprocessing on the data such as stop words removal. We included only those bigrams in the feature vector that occurred more than three times in the two corpora (fraud, no-fraud). We used bigrams as they can be easily extracted and it is possible to capture more context with bigrams. For instance, even though the classifier knows that keyword ‘allege’ is a fraud indicator, it is difficult to conclude that a statement containing ‘allege’ is a fraudulent one. This is just one of the many examples that are confusing to the classifier when we used a keyword-based approach to classification.

6.2 FEATURE SELECTION AND CLASSIFIER RESULTS

Feature selection is typically performed by selecting a subset of features from the original set of features. In this study, we used a forward feature stepwise selection approach to feature selection and incrementally added a feature, one at a time in the feature space (see Chapter 3 for more information on the feature selection approach used in this study). This way we were able to understand the role of different features on the classifier
performance and construct a feature set that was most relevant for fraud detection and
detection of different stages of fraud.

In the text categorization and document classification literatures, several feature selection
measures have been explored such as information gain, mutual information, document
frequency and chi-square test. Information gain ratio is a popular measure in the field of
machine learning to determine the degree of contribution made by a feature in accurately
predicting the distribution of classes. It measures the number of bits of information
received for category prediction by knowing the presence or absence of a feature in a
document.

We used chi-square method to select features that show statistically significant
differences between the positive and negative documents. Chi-square feature selection
has been shown to not only reduce the feature space effectively by reducing the noise
introduced in the classifier, but also to improve performance of the classifier at the same
time. A 2x2 contingency (cross-classification) table was constructed for each feature and
the p-value was computed using chi-square. The p-value was used to decide whether or
not the null hypothesis can be rejected. Features with p-values less than the ‘alpha’ value
for the chosen level of confidence ($\alpha = .05$) were used as input to the classification
algorithm. This statistic also indicated if the likelihood of detecting fraud is related or not
related to this feature.

**6.2.1 TOP RANKING FEATURES**

Using the chi-square test of significance, we were able to eliminate those features for
which we were not able to reject the null hypothesis. The features that did not contribute
to classifier performance (with the highest p-value) such that their inclusion made no difference in the classifier accuracy were also eliminated from the feature space. Thus, our final set of selected features included only those features that played a role in increasing the overall accuracy of the fraud classifier. Tables 6.6 and 6.7 provide a ranking of the top ten features for recognizing fraudulent and non-fraudulent annual reports and for recognizing different stages of fraud, respectively.

**TABLE 6.6**
*Top Ten Features for Detecting Fraudulent and Non-Fraudulent Annual Reports*

<table>
<thead>
<tr>
<th>Rank</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Percentage of Passive Voice Sentences</td>
</tr>
<tr>
<td>2</td>
<td>Percentage of Active Voice Sentences</td>
</tr>
<tr>
<td>3</td>
<td>Standard Deviation of Sentence Lengths</td>
</tr>
<tr>
<td>4</td>
<td>Readability Index</td>
</tr>
<tr>
<td>5</td>
<td>Scaled Frequency of Uncertainty Markers</td>
</tr>
<tr>
<td>6</td>
<td>Percentage of Sentences Beginning with Subordinating Conjunction</td>
</tr>
<tr>
<td>7</td>
<td>Type-token Ratio</td>
</tr>
<tr>
<td>8</td>
<td>Scaled Frequency of Proper Nouns</td>
</tr>
<tr>
<td>9</td>
<td>Percentage of “To be” Verbs</td>
</tr>
<tr>
<td>10</td>
<td>TFIDF Weighted Tokens</td>
</tr>
</tbody>
</table>

**TABLE 6.7**
*Top Ten Features for Detecting Stages of Fraud*  
(‘Pre-Fraud,’ ‘Advanced-Fraud’)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Readability Index</td>
</tr>
<tr>
<td>2</td>
<td>Percentage of Passive Voice Sentences</td>
</tr>
<tr>
<td>3</td>
<td>Percentage of Active Voice Sentences</td>
</tr>
<tr>
<td>4</td>
<td>Standard Deviation of Sentence Lengths</td>
</tr>
<tr>
<td>5</td>
<td>Scaled Frequency of Uncertainty Markers</td>
</tr>
<tr>
<td>6</td>
<td>Type-token Ratio</td>
</tr>
<tr>
<td>7</td>
<td>Percentage of Words that Belong to ‘Positive’ Tone Category</td>
</tr>
<tr>
<td>8</td>
<td>Standard Deviation of Word Lengths</td>
</tr>
<tr>
<td>9</td>
<td>Percentage of Sentences Beginning with Subordinating Conjunction</td>
</tr>
<tr>
<td>10</td>
<td>Scaled Frequency of Proper Nouns</td>
</tr>
</tbody>
</table>
We ran classifier experiments with single features as well as combinations of the features contained in the four feature sets described earlier. Our classification results showed that when we used keywords as a single feature, the classifier performance did not improve over what we had with our baseline results. In addition, when keywords were used in combination with other features, their inclusion did not make any tangible improvement in the classifier performance. This may be due to the fact that a keywords-based approach, which seems to work effectively for tasks such as topic classification, has limited applicability to the more demanding task of fraud classification. By using bigrams, we were able to improve the classifier accuracy by about 2%. Our findings were consistent with some of the earlier studies (Dulaney, 1982; Colwell et al., 2002; Zhou et al., 2002) in relation to type-token ratio. Earlier researchers noted that higher type-token ratio was associated with deception. We also discovered that fraudulent annual reports had a higher type-token ratio. Conversely, non-fraudulent annual reports were interpreted to be more truthful as indicated by lower type-token ration.

### 6.2.2 Fraud Model Results

Our fraud classifier results with the highest ranked features yielded an accuracy of 89.51%, which was much higher than our baseline results of 71.67% that we obtained using the “bag of words” approach. Table 6.8 presents average scores of classifier accuracy over ten-fold cross-validation. The confusion matrix results and detailed accuracy results for each of the three datasets are presented in Tables 6.9 through 6.14.
### TABLE 6.8

*Average Classifier Accuracy for the Three Datasets with most Useful Features*

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Fraud Detection Version 1</th>
<th>Fraud Detection Version 2</th>
<th>Detection of Stages of Fraud</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fraud</td>
<td>NoFraud</td>
<td>Fraud</td>
</tr>
<tr>
<td># Of Documents</td>
<td>405</td>
<td>622</td>
<td>405</td>
</tr>
<tr>
<td>Average Classifier Accuracy</td>
<td>89.51%</td>
<td>89.04%</td>
<td>87.98%</td>
</tr>
</tbody>
</table>

### TABLE 6.9

*Confusion Matrix Results for Fraud Version 1 Dataset*

<table>
<thead>
<tr>
<th>Class</th>
<th>Fraud (Predicted)</th>
<th>NoFraud (Predicted)</th>
<th>Total (Actual)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraud</td>
<td>364</td>
<td>41</td>
<td>405</td>
</tr>
<tr>
<td>NoFraud</td>
<td>66</td>
<td>556</td>
<td>622</td>
</tr>
<tr>
<td>Total (Predicted)</td>
<td>430</td>
<td>597</td>
<td>1027</td>
</tr>
</tbody>
</table>

### TABLE 6.10

*Detailed Accuracy Results for Fraud Version 1 Dataset*

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraud</td>
<td>0.899</td>
<td>0.106</td>
<td>0.847</td>
<td>0.899</td>
<td>0.872</td>
</tr>
<tr>
<td>NoFraud</td>
<td>0.894</td>
<td>0.101</td>
<td>0.931</td>
<td>0.894</td>
<td>0.912</td>
</tr>
</tbody>
</table>
**TABLE 6.11**
*Confusion Matrix Results for Fraud Version 2 Dataset*

<table>
<thead>
<tr>
<th>Class</th>
<th>Fraud (Predicted)</th>
<th>NoFraud (Predicted)</th>
<th>Total (Actual)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraud</td>
<td>362</td>
<td>43</td>
<td>405</td>
</tr>
<tr>
<td>NoFraud</td>
<td>107</td>
<td>863</td>
<td>970</td>
</tr>
<tr>
<td>Total (Predicted)</td>
<td>469</td>
<td>906</td>
<td>1375</td>
</tr>
</tbody>
</table>

**TABLE 6.12**
*Detailed Accuracy Results for Fraud Version 2 Dataset*

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraud</td>
<td>0.894</td>
<td>0.110</td>
<td>0.772</td>
<td>0.894</td>
<td>0.828</td>
</tr>
<tr>
<td>NoFraud</td>
<td>0.890</td>
<td>0.106</td>
<td>0.952</td>
<td>0.890</td>
<td>0.920</td>
</tr>
</tbody>
</table>

**TABLE 6.13**
*Confusion Matrix Results for Detection of Stages of Fraud Dataset*

<table>
<thead>
<tr>
<th>Class</th>
<th>Adv (Predicted)</th>
<th>Pre (Predicted)</th>
<th>Total (Actual)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adv</td>
<td>358</td>
<td>47</td>
<td>405</td>
</tr>
<tr>
<td>Pre</td>
<td>27</td>
<td>181</td>
<td>208</td>
</tr>
<tr>
<td>Total (Predicted)</td>
<td>385</td>
<td>228</td>
<td>613</td>
</tr>
</tbody>
</table>

**TABLE 6.14**
*Detailed Accuracy Results for Detection of Stages of Fraud Dataset*

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adv</td>
<td>0.884</td>
<td>0.130</td>
<td>0.930</td>
<td>0.884</td>
<td>0.906</td>
</tr>
<tr>
<td>Pre</td>
<td>0.870</td>
<td>0.116</td>
<td>0.794</td>
<td>0.870</td>
<td>0.830</td>
</tr>
</tbody>
</table>
6.2.3 DISCUSSION

These results support our claim that annual reports contain linguistic cues that can be exploited to proactively detect fraud. Furthermore, these results suggest that the subset of features that we have selected for fraud detection can be used successfully to distinguish fraudulent annual reports from non-fraudulent annual reports 89.51% of the time. Similarly, the subset of features that we have selected for detection of stages of fraud can be used successfully to distinguish early symptoms of fraud from advanced stages of fraud 87.98% of the time.

For the best score of 89.51 percent, these results indicate that the FP rate of 0.101 for the no-fraud class is lower than the FP rate of 0.106 for the fraud class. This means that the classifier missed only 10.1 percent of the fraudulent annual reports (type I error) and misclassified them as non-fraudulent annual reports, whereas the classifier misclassified non-fraudulent annual reports as fraudulent annual reports (type II error) 10.6 percent of the time.

These results also indicate that the TP rate of 0.899 for the fraud class is higher than the TP rate of 0.894 for the no-fraud class. Since a classifier is considered superior to another if it has a higher TP rate and a lower FP rate, we can conclude that the performance of the classifier with these top ten features is far superior to its performance with baseline features. For the no-fraud class, the TP rate of 0.894 is lower than the TP rate of 0.913, which was obtained with baseline features. However, it should be noted that this is a more desirable situation for the no-fraud class as its FP rate has tremendously gone down from 0.585 obtained with baseline features to 0.101,
As observed in Table 6.10, the recall rate for the minority class ‘fraud’ has gone up to 0.899 from 0.415 achieved with baseline experiments and its predictive accuracy has also increased to 0.847 from 0.757. This indicates that the classifier was able to overcome the class imbalance problem as it converged to higher levels of learning. Thus, our intuition was correct in the respect that as the classifier was trained with more sophisticated features, the learning accuracy of the classifier increased even for the minority class ‘fraud.’ When we performed our baseline experiments with the bag-of-words approach, precision was low as there were many unnecessary features.

When these results for fraud detection were compared to the random baseline, we noted that the classifier was able to beat the random baseline by a much wider margin. The random baseline would yield an accuracy of 60.56 percent simply by classifying all the documents to the largest class (no-fraud).

We also ran our experiments with different parameters and kernels such as the radial basis function kernel for the SVM but selection of other kernels had very little impact on improving the fraud classification accuracy. Our best results were achieved using a linear kernel with default parameters. In addition, we tested our fraud model built with top ranked features on unseen data. The unseen data consisted of 8-Ks of fraudulent and non-fraudulent companies. The ten-fold cross-validation accuracy of the fraud classifier was 85.32%. As noted earlier, the reported results provide clear evidence that the classifier, built using linguistic features, offers high performance and can be extended to a large scale of data.
6.3 SUMMARY

In this chapter we discussed the features that we used to examine the presentation style and content of the annual reports for detecting fraud and stages of fraud. We also described the feature extraction process and how we converted these raw scores into normalized scores. In addition, we provided an overview of the feature selection process that we implemented. We also presented subsets of features that we found most useful for recognizing fraudulent and non-fraudulent annual reports and different stages of fraud. Our fraud classifier results with the highest ranked features yielded an accuracy of 89.51 percent, which is much higher than our baseline results of 71.67 percent that we obtained using the “bag of words” approach. In addition, our classifier results for detection of stages of fraud with the highest ranked features yielded an accuracy of 87.98 percent, which is much higher than our baseline results of 65.81 percent. These results support our claim that fraud can be proactively detected by doing linguistic analyses of the textual content of annual reports.
CHAPTER 7
HYPOTHESES TESTING AND EVALUATION

This chapter describes the results of the hypothesis tests and discusses their implications. Section 7.1 presents results of individual hypothesis testing using the chi-square test of significance. Section 7.2 summarizes the key results of the hypothesis tests.

7.1 RESULTS OF HYPOTHESES TESTING

\( H1: \) The greater the use of complex sentential structures in the qualitative content of a company’s annual report, the greater the likelihood that there will be fraud.

Our first hypothesis, H1, was based on the notion that companies committing fraud tend to present their annual reports in a convoluted style consciously employing means to obscure real information that might expose fraud. This hypothesis rests on the idea that fraudulent annual reports make greater use of complex sentential structures and thus exhibit high levels of both lexical and structural ambiguity. Conversely, clear succinct language is employed to convey information more clearly, which would require short and direct sentential constructs.

Related to hypothesis H1, we also defined and tested a sub hypothesis, H1a, as follows:

\( H1a: \) The more difficult it is to read and understand a company’s annual report, the greater the likelihood that there is fraud.
Hypothesis H1a is of interest as it is believed that, in cases of fraud, companies deliberately employ tools to make annual reports difficult to read and comprehend. Therefore, fraudulent annual reports show greater use of longer sentences and difficult words. Conversely, to make reports easier to understand, management makes precise, molecular statements by making greater use of shorter sentences with simple words so that the readability of such reports is improved.

For testing H1, the null hypothesis that the count of complex sentential structures as indicated by an ambiguity index obtained through DICTION, the frequency of sentences beginning with a subordinating conjunction, the frequency of different word types, and the frequency of function words is similar in both fraudulent and non-fraudulent annual reports was tested. We examined the frequency of function words, as these words are typically the most frequent words. Prior studies in linguistics have shown that documents containing more frequent words are easy to read when everything else is same (Graham et al., 2005). For testing H1a, the null hypothesis that the readability index is similar in both fraudulent and non-fraudulent annual reports was tested.

The results of chi-square tests for hypotheses H1 and H1a indicated that we could reject the null hypotheses and conclude that the variation between the fraudulent and non-fraudulent annual reports is due to systematic alterations and not due to chance (p < 0.001 and < 0.05 for H1 and H1a, respectively).
H2: The greater the use of negative words in a company’s annual report, the greater the likelihood that there is fraud.

Hypothesis H2 is of interest since it is believed that management consciously chooses selective accounting language to conceal fraud. The annual report typically reflects a company’s strategy and when things are not going well, its numbers do not change but qualitative content changes. For H2, the null hypothesis, that negative and positive category words are used similarly in the qualitative content of fraudulent and non-fraudulent annual reports, was tested, i.e., the null hypothesis stated that there exists no relationship between the polarity of tone and the outcome of fraud. Tables 7.1 and 7.2 show contingency tables for both versions of the fraud detection dataset. Table 7.3 presents chi-square results of testing the null hypothesis for H2.

**TABLE 7.1**

*Contingency Table for Testing a Fraud Hypothesis Relating to Tone: Version 1*

<table>
<thead>
<tr>
<th>Features Category</th>
<th>Negative Words</th>
<th>Positive Words</th>
<th>Other Words (Residual)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraudulent Annual Reports</td>
<td>57689</td>
<td>67902</td>
<td>9813656</td>
</tr>
<tr>
<td>Non-Fraudulent Annual Reports</td>
<td>81529</td>
<td>97312</td>
<td>14186253</td>
</tr>
</tbody>
</table>
### TABLE 7.2

*Contingency Table for Testing a Fraud Hypothesis Relating to Tone: Version 2*

<table>
<thead>
<tr>
<th>Features Category</th>
<th>Negative Words</th>
<th>Positive Words</th>
<th>Other Words (Residual)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraudulent Annual Reports</td>
<td>57689</td>
<td>67902</td>
<td>9813656</td>
</tr>
<tr>
<td>Non-Fraudulent Annual Reports of Mixed Companies</td>
<td>139851</td>
<td>162507</td>
<td>22575512</td>
</tr>
</tbody>
</table>

### TABLE 7.3

*Chi-Square Results for the Tone Category Words*

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Degrees of Freedom</th>
<th>Chi-Square</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraud Detection Version 1</td>
<td>2</td>
<td>20.04</td>
<td>p ≤ 0.001</td>
</tr>
<tr>
<td>Fraud Detection Version 2</td>
<td>2</td>
<td>184.88</td>
<td>p ≤ 0.001</td>
</tr>
</tbody>
</table>

The results of chi-square tests for hypothesis H2 indicated that we could reject the null hypothesis and conclude that the variation in the use of positive and negative tone category words between the fraudulent and non-fraudulent annual reports is significant. However, it must be noted that when we look at the relative distribution of negative and positive category words between the two corpora for the first version, there is not much difference. For example, negative and positive words have similar percentages [45.93% (negative) and 54.07% (positive) out of total of 125,591 tone category words in the fraud corpus compared to 45.59% (negative) and 54.41% (positive) out of total of 178,841 tone...
category words in the no-fraud corpus]. Nevertheless, the normalized counts of the negative and the positive category words showed that the fraud corpus has 5,804.16 negative and 6,831.70 positive instances per 1,000,000 words, compared to 5,675.49 negative and 6,774.19 positive instances in the no-fraud corpus.

The analysis of these results suggests that it might be useful to look at the context of neighboring words and sentences before we draw any conclusions on the presence of target tone category words. The omission of context leads to a noisy measure of tone since there is only limited evidence to support the claim that positive category tone words denote positive connotations and negative category tone words denote negative connotations. Thus, counting tone category words independent of the context of usage may be counterintuitive especially for the task of fraud detection. For example, consider the two sentences mentioned below:

(1) The company has not been alleged in any lawsuits.

(2) The company has been alleged in a number of lawsuits.

If negative tone category words are ‘alleged’ and ‘lawsuits,’ then count of these words in two sentences (1) and (2) is similar. On the basis of frequency count of tone category words, it is difficult to conclude that tone is different for two sentences. However, if we look at the context of these tone category words, then sentence (1) has a positive connotation whereas sentence (2) has a negative connotation.
H3: The greater the use of passive voice in a company’s annual report, the greater the likelihood that there is fraud.

Hypothesis H3 is of interest as it is believed that, in cases of fraud, management consciously tries to shift responsibility away from itself and makes greater use of passive voice rather than active voice. Conversely, in cases of no-fraud, in order to take credit for positive outcomes and attribute these outcomes to its own actions, management will use active voice sentences. For H3, the null hypothesis that there is no difference in the use of voice in the qualitative content of fraudulent and non-fraudulent annual reports was tested. Tables 7.4 and 7.5 show contingency tables for both versions of the fraud detection dataset. Table 7.6 presents chi-square results of testing the null hypothesis for H3.

**TABLE 7.4**

*Contingency Table for Testing a Fraud Hypothesis Relating to Voice: Version 1*

<table>
<thead>
<tr>
<th>Features Category</th>
<th>Passive Voice</th>
<th>Active Voice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraudulent Annual Reports</td>
<td>98379</td>
<td>89765</td>
</tr>
<tr>
<td>Non-Fraudulent Annual Reports</td>
<td>299161</td>
<td>300898</td>
</tr>
</tbody>
</table>
TABLE 7.5
Contingency Table for Testing a Fraud Hypothesis Relating to Voice: Version 2

<table>
<thead>
<tr>
<th>Features</th>
<th>Passive Voice</th>
<th>Active Voice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraudulent Annual Reports</td>
<td>98379</td>
<td>89765</td>
</tr>
<tr>
<td>Non-Fraudulent Annual Reports of Mixed Companies</td>
<td>466911</td>
<td>467812</td>
</tr>
</tbody>
</table>

TABLE 7.6
Chi-Square Results for Use of Passive Voice and Active Voice Sentences

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Degrees of Freedom</th>
<th>Chi-Square</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraud Detection Version 1</td>
<td>1</td>
<td>339.44</td>
<td>p ≤ 0.001</td>
</tr>
<tr>
<td>Fraud Detection Version 2</td>
<td>1</td>
<td>342.29</td>
<td>p ≤ 0.001</td>
</tr>
</tbody>
</table>

The results of chi-square tests for hypothesis H3 indicated that we could reject the null hypotheses with very high confidence and conclude that the variation in the use of passive and active voice sentences between the fraudulent and non-fraudulent annual reports is due to systematic alterations and not due to chance (p < 0.001). A closer examination of the distribution of passive/active voice construction in the two corpora shows a high incidence of passive instead of active voice sentences in the fraud corpus. This explanation is supported by the predominance of the passive voice sentences in the
fraudulent annual reports as opposed to the non-fraudulent annual reports, which showed high figures for active voice sentences.

*H4: The greater the use of uncertainty markers in a company’s annual report, the greater the likelihood that there is fraud.*

Hypothesis H4 is of interest as it is believed that, in cases of fraud, management deliberately employs more uncertainty markers to make reports uncertain. The management may rationalize this act by arguing that it is better to introduce uncertainty than to make false statements. Conversely, in cases of no-fraud, to provide a definite and clear picture, management will avoid using uncertainty markers. For H4, the null hypothesis, that the use of uncertainty markers is similar in the qualitative content of both fraudulent and non-fraudulent annual reports was tested. Tables 7.7 and 7.8 show contingency tables for both versions of the fraud detection dataset. Table 7.9 presents chi-square results of testing the null hypothesis for H4.

**TABLE 7.7**  
*Contingency Table for Testing a Fraud Hypothesis Relating to Uncertainty Markers: Version 1*

<table>
<thead>
<tr>
<th>Features</th>
<th>Uncertainty Markers</th>
<th>Other Words (Residual)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraudulent Annual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reports</td>
<td>115722</td>
<td>9823525</td>
</tr>
<tr>
<td>Non-Fraudulent</td>
<td>157934</td>
<td>14207160</td>
</tr>
<tr>
<td>Annual Reports</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The results of chi-square tests for hypothesis H4 indicated that we could reject the null hypotheses with very high confidence and conclude that the variation in the use of uncertainty markers between the fraudulent and non-fraudulent annual reports is due to systematic alterations and not due to chance ($p < 0.001$). Furthermore, the normalized counts of uncertainty markers showed that the fraud corpus has 11,642.93 instances per 1,000,000 words, compared to 10,994.29 instances in the no-fraud corpus for the first version of the dataset.

**TABLE 7.8**

*Contingency Table for Testing a Fraud Hypothesis Relating to Uncertainty Markers: Version 2*

<table>
<thead>
<tr>
<th>Features Category</th>
<th>Uncertainty Markers</th>
<th>Other Words (Residual)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraudulent Annual Reports</td>
<td>115722</td>
<td>9823525</td>
</tr>
<tr>
<td>Non-Fraudulent Annual Reports of Mixed Companies</td>
<td>252799</td>
<td>22625071</td>
</tr>
</tbody>
</table>

**TABLE 7.9**

*Chi-Square Results for Use of Uncertainty Markers*

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Degrees of Freedom</th>
<th>Chi-Square</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraud Detection Version 1</td>
<td>1</td>
<td>222.02</td>
<td>$p \leq 0.001$</td>
</tr>
<tr>
<td>Fraud Detection Version 2</td>
<td>1</td>
<td>219.44</td>
<td>$p \leq 0.001$</td>
</tr>
</tbody>
</table>
7.2 SUMMARY

The results of the hypothesis tests presented in this chapter were in line with our expectations, as we were able to reject the null hypotheses corresponding to each of our hypotheses. The chi-square tests of significance supported our hypotheses for sentence complexity, readability, tone, use of voice, and uncertainty markers. This implies that the features such as the use of active and passive voice sentences, use of uncertainty markers, readability, ambiguity, frequency of sentences beginning with a subordinating conjunction, and use of different word types can be used to differentiate the presentation style of the fraudulent and non-fraudulent annual reports. This was evident by the differences in the distributions of these features between fraudulent and non-fraudulent annual reports.
CHAPTER 8
SUMMARY AND CONCLUSION

This chapter provides a brief overview of main findings, implications, limitations and directions for future research. Section 8.1 provides a summary of the research and the main findings. Section 8.2 discusses implications of these findings for both academicians and practitioners. Section 8.3 presents some inherent limitations. Section 8.4 outlines directions for future research. Finally, section 8.5 presents concluding remarks.

8.1 SUMMARY AND MAIN FINDINGS

In this dissertation we presented a methodology to proactively detect fraud by examining the qualitative content of annual reports using natural language processing tools. We investigated both the verbal content and the presentation style of the annual reports and provided evidence that linguistic features are useful in distinguishing fraudulent annual reports from non-fraudulent annual reports. Our analysis was based upon a sample of 126 companies identified with financial statement fraud and 622 no-fraud companies matched by two-digit SIC code, size and year.

In order to avoid recognizing the company’s style rather than the presence of fraud, we also performed a comparative analysis on two no-fraud datasets to evaluate the effects of a company’s style on the performance of fraud detection model. For this, we created two versions of a no-fraud dataset and each was paired with a fraud dataset. The first version
of the no-fraud dataset consisted of 10-Ks of only no-fraud companies. The second version of the no-fraud dataset included 10-Ks for non-fraudulent years of fraud companies that were outside the pre-fraud, fraud, and post-fraud periods in addition to the 10-Ks of no-fraud companies.

Using the methodology described in this study, we ran all our experiments on three datasets. In the third dataset, we examined the changes in the qualitative content of annual reports of fraud companies for pre-fraud, fraud, and post-fraud periods to detect stages of fraud in order to distinguish early symptoms of fraud from advanced symptoms of fraud.

Our initial baseline results with a Naïve Bayes classifier, using a ‘bag of words’ approach were modest, correctly classifying about 56.75 percent (best among the four different combinations of features). However, when we used Support Vector Machines (SVM) as the main classifier, the fraud classification accuracy even with baseline features increased to 71.67 percent. Inspired by encouraging baselines results, we ran our classifier experiments with four feature sets to detect fraud and stages of fraud. The first feature set consisted of eight simple stylistic surface features such as the average and standard deviation of the lengths of sentences. The second feature set consisted of other style features such as voice, readability index, tone, and uncertainty markers. The third feature set consisted of deeper stylistic surface features that required deeper linguistic analysis such as verbal categories (auxiliary verbs, “to be” verbs), type-token ratio, and different sentence openers. The fourth feature set consisted of content related features such as keywords, bigrams, and TFIDF-weighted tokens.
The pre-selection of these features was inspired by our informed reasoning and domain knowledge and rested on the speculation that the qualitative content of annual reports manifests linguistic cues that can be harvested for predicting fraud. As the fraud classifier converged to higher levels of accuracy, we isolated features that contributed the most to the classifier performance.

Our fraud classifier results with the highest ranked features yielded an accuracy of 89.51%, which was much higher than our baseline results of 71.67% that we obtained using a “bag of words” approach with SVM. The classifier results for detection of stages of fraud with the highest ranked features yielded an accuracy of 87.98 percent, which was much higher than our baseline results of 65.81 percent. These results support our claim that annual reports contain linguistic cues that can be exploited to proactively detect fraud.

To our knowledge, no prior studies have used linguistic features to examine the qualitative content of annual reports in the context of fraud detection. The major thrust of prior studies has been on empirically examining the relationship between fraudulent financial reporting and quantitative information such as ratios and quantitative indicators such as the composition of boards of directors, insider trading, auditor rotation, or financial restatements. Even though researchers have examined qualitative components of annual reports, very few studies have used the entire qualitative content to predict fraud. Building on prior studies, the present study bridges this research gap by using an alternative methodology for examining and detecting fraud.
8.2 IMPLICATIONS

From an academic perspective, our research findings contribute primarily to two lines of research. First, we identify linguistic variables that deserve further research as potential new signals of fraud. The linguistic differences reported in fraudulent and non-fraudulent annual reports are not meant to oversimplify issues of detecting fraud, but to provide insight and understanding into the ways the companies portray themselves that require further investigation. Our work in the area of fraud detection using linguistic analysis has opened up many possibilities by adding another dimension to corporate fraud research. For example, motivated by our research findings, more sophisticated linguistic features can be explored to see if their inclusion can further improve the performance of fraud detection model built in this study.

Second, we provide additional usefulness of qualitative content of annual reports for detecting fraud. Extensive research has been done on the analysis of financial data; however, there is scant research on the analysis of text in financial statements to detect fraud.

Our approach is unique as it combines two very different fields, namely accounting and linguistics for a very useful application, to predict early warning signs of forthcoming accounting problems in potential fraudulent companies, which can be of interest to practitioners such as auditors, analysts and fraud examiners. For practitioners as well as academicians, our fraud model suggests that adding a linguistic dimension to fraud examination can increase the accuracy of fraud detection. Our fraud model can also help them in finding early symptoms of fraud.
Furthermore, our findings should be of interest to practitioners for the purpose of fraud risk assessment. Our research methodology can aid auditors in fraud risk assessment by classifying annual report as fraudulent, which would indicate an increased risk of fraud, thus signaling the auditor, of companies that are at higher risk for fraud. Auditors face adverse legal and regulatory consequences if fraud goes undetected. As a consequence, it is crucial for auditors to accurately do fraud risk assessment as incorrect assessment of fraud risk may result in failure to detect fraudulent reporting, thereby increasing the auditor’s exposure.

Our fraud model can also help regulators such as the Securities and Exchange Commission (SEC) by flagging potential fraudulent companies and SEC can pursue these suspicious companies for further investigation. While conventional fraud models can also predict the targets of SEC investigation, their results have not been very promising due to poor prediction accuracy. The conventional fraud models were built using quantitative financial metrics that have limited ability to detect and/or predict fraud (Kaminski et al., 2004).

The fraud model built in this study is easily portable to other datasets. For example, when we tested our fraud model on unseen data, which consisted of 8-Ks of fraudulent and non-fraudulent companies, the fraud classifier achieved an accuracy of 85.32 percent. Our empirical results also indicate that SVM is well suited for capturing patterns and for detection problems in accounting and auditing domains.
8.3 LIMITATIONS

Our fraud dataset consists of fraud companies that have documented evidence of fraud. These companies might not be representative of all companies that have committed fraud. This is due to the fact that companies that commit fraud are less forthcoming about this information (Higson, 1999). As a result, fraud may exist in such companies without it being publicly revealed or discovered. For example, a survey by KPMG (1998) found that over one third of fraud cases were discovered by accident. Like most other fraud studies, companies with undetected frauds are not included in the fraud sample. Furthermore, the fraud sample was limited to publicly listed companies and did not include private companies. Another limitation relates to the sample of no-fraud companies. For the same reason mentioned above (i.e., companies committing fraud are not forthcoming with this information), no-fraud companies may also include companies where fraud had occurred but it has not been publicly discovered.

Another limitation relates to the fact that supervised learning algorithms cannot discover a novel feature unless it is either learned from the training dataset or defined by a user. In addition, if a dataset is imbalanced, then SVMs tend to produce a less effective classification boundary skewed to the minority class. When there are too few positive examples, SVMs may totally fail, as there is insufficient evidence for statistical learning. It is not clear as to how large a dataset should be or how many examples should be provided for each class.

Another common issue with supervised classification models is the issue of generalization. The basic assumption of supervised classification is that both training and
test datasets belong to the same population. In order to determine the degree of
generalization, it is important to test the classifier on a population that is different from
the training and test datasets upon which it was built and tested. However, this problem
of generalization cannot be easily solved due to the limited availability of benchmark
collections that are suitable to one’s domain problem (Wulfekuhler and Punch, 1997;
Yang, 2001). As a result, supervised classification models remain strongly conditioned by
the characteristics of the population of the training set. Furthermore, for fraud detection,
it is difficult to compare the results of different researchers, as there is no common
dataset or methodology that has been used to build fraud detection models.

8.4 FUTURE RESEARCH

Our reported results showed that tone was among the top ten features for detecting stages
of fraud, even though it was not present among the top ten features for detecting fraud.
One possible explanation is that context of the positive and negative words may have an
effect on tone analysis. We might yield better results if we also look at the context of
these words by examining a small number of words around the target positive and
negative tone words. Thus, examination of alternative methodologies for capturing tone
deserves further research. It may be worthwhile to recode the tone category words as that
could allow better identification and analysis of tone as a feature. In addition, use of a
system that uses context to assign tone category words will be useful. This issue is left for
future study.

As technology is facilitating on-line access to information, the role of text analytics and
information retrieval in the accounting domain will progress from simple applications to

176
more sophisticated applications. One direction for future research in the area of fraud
detection can be to create a fraud ontology that not only covers corporate fraud but also
other types of frauds. The relationships between different types of fraud will help in
finding fraud keywords that can be tested for their role in fraud detection using the
methodology presented in this research. Development of a fraud ontology can also
benefit from thematic analyses of accounting narratives to identify salient content
categories in the context of fraud. This will also help in the construction of a fraud
taxonomy. More applications using Natural Language Processing techniques for
information extraction and classification are required to successfully extract and organize
information in terms of extraction of fraud synonyms or related terms and their automatic
classification into functional hierarchies. In addition, it will be beneficial to investigate
how new words in the accounting domain are formed as this kind of analysis will help in
formulating rules of word formation, and in the future contribute towards maintaining
and extending a fraud ontology/taxonomy in response to changes in the accounting
domain.

8.5 CONCLUSION

In this research, we presented a methodology that involved linguistic analysis of the
textual content of annual reports for detecting fraud. Linguistic cues not only helped us in
interacting with the text and but also allowed us to look beyond the content of the annual
reports. By doing both stylistic analysis and content analysis of these annual reports, we
were able to build a fraud detection model that is competitive with the leading fraud
detection models and achieves very good results in terms of precision and recall. We
were able to improve the predictive accuracy of our fraud detection model from baseline
results of 71.67 percent accuracy using a “bag of words” approach to 89.51 percent accuracy when we incorporated linguistically motivated features inspired by our informed reasoning and domain knowledge. Our results showed that by adding a linguistic dimension to fraud examination, accuracy of fraud detection could be substantially improved.

A possible explanation for the greater predictive accuracy of the fraud model built with the linguistic features capturing the qualitative content and style of the annual reports is that the quantitative financial numbers contain redundant information that does not change when a company is committing a fraud but the writing style and the presentation style employed to communicate financial information changes. For example, most of the fraudulent companies were effective at hiding their fraudulent activities with the illusion of strong financial performance to forestall shareholder investigation into management actions.

Our work is among the first to use the qualitative content of annual reports to detect fraud. Indeed, our study samples the whole text of the annual reports instead of selected passages and hence it is more robust. Our reported results are based on observations of companies that were in different industries and for different time periods, and hence the results can be generalized to other industries and time periods. Unlike some of the earlier fraud models that were based on synthetic data, our fraud model is based on real data. Furthermore, our reported results are more credible as many of the earlier studies on fraud detection were based on much smaller datasets.
The results of our study suggest that the qualitative narrative content of annual reports contains information that is useful for detecting fraud that is not accurately captured by financial numbers. We found systematic differences in communication and writing style of fraudulent annual reports such as fraudulent annual reports contained more passive voice sentences, used more uncertainty markers, had a higher type-token ratio (lexical variety), and were more difficult to read and comprehend than non-fraudulent annual reports.
REFERENCES


49. Donaldson, T. (2005), “Corporate Fraud on Tiral: What have we learned?” Knowledge@Wharton (http://knowledge.wharton.upenn.edu/article.cfm?articleid=1131)


85. Institute of Internal Auditors (1985), SIAS 3- “Deterrence, Detection, Investigation, and Reporting of Fraud”, The Institute of Internal Auditors, Altamonte Springs, FL.


175. Yu, H., Yang, J., and Han, J. (2003), “Classifying Large Data Sets Using SVMs with Hierarchical Clusters”, Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.


