

PREDICTING FINANCIAL FRADULUENT STATEMENTS USING SEMANTICS

PROJECT REPORT

Submitted by

B.Archana	(10Z305)
R.K.Manisha	(10Z322)
S.Sasikala	(10Z340)
V.Subhaprbha	(10Z349)
A.Lalitha	(11Z462)

Dissertation submitted in partial fulfillment of the requirements for the degree of

BACHELOR OF ENGINEERING

Branch: COMPUTER SCIENCE AND ENGINEERING
of Anna University



April 2014

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

PSG COLLEGE OF TECHNOLOGY
(Autonomous Institution)

COIMBATORE-641 004

PSG COLLEGE OF TECHNOLOGY

(Autonomous Institution)

COIMBATORE – 641 004

PREDICTING FINANCIAL FRADULUENT STATEMENTS USING SEMANTICS

Bonafide record of work done by

B.Archana	(10Z305)
R.K.Manisha	(10Z322)
S.Sasikala	(10Z340)
V.Subhaprba	(10Z349)
A.Lalitha	(11Z462)

Dissertation submitted in partial fulfillment of the requirements for the degree of

BACHELOR OF ENGINEERING

Branch: **COMPUTER SCIENCE AND ENGINEERING**

of Anna University

APRIL 2014

.....
Mrs. C. Kavitha

Faculty guide

.....
Dr.R.Venkatesan

Head of the Department

Certified that the candidate was examined in the viva-voce examination held on

.....
(Internal Examiner)

.....
(External Examiner)

CONTENTS

CHAPTER	Page No.
Acknowledgement.....	(i)
Synopsis.....	(ii)
List of figures.....	(iii)
List of tables.....	(iv)
1. INTRODUCTION.....	1
1.1. Motivation	1
1.2. Objective	2
1.3. Need for the project	2
1.4. Advantage	2
2. LITERATURE SURVEY.....	3
2.1 Predicting fraudulent financial statements with machine learning techniques	3
2.2 Detecting financial statement fraud	3
2.3. Prevention and detection of financial statement fraud-an implementation of data mining framework	4
2.4 Detection of fraudulent financial statements through the use of data mining techniques	4
2.5 Multi-instance learning for predicting fraudulent financial statement	4
2.6 A review of financial accounting fraud detection based on data mining Techniques	5
3. REQUIREMENTS.....	6
3.1. Requirement Analysis	6
3.2. Feasibility Analysis	7
3.2.1 Technical Feasibility	7
3.2.2 Financial Feasibility	7
3.2.3 Schedule Feasibility	7
3.2.4 Resource Feasibility	7

4. SYSTEM DESIGN.....	8
4.1. Activity Diagram	8
4.2. Usecase Diagram	9
5. SYSTEM IMPLEMENTATION.....	11
5.1. Report Processing	11
5.2. Parameters	12
5.3. Parsing	12
5.4. Dictionary Construction	13
5.5. Score card Preparation	14
5.6. Feature Selection	15
5.6.1 T-Test Evaluation	15
5.7. Associative Classification	15
5.7.1 Steps in Associative Classification Rule Mining	16
6. EXPERIMENTAL RESULTS.....	17
6.1. Score card formation	17
6.2. Feature Selection	18
6.3. Associative Classification	19
6.4. Summary of Dataset	19
6.5. Accuracy of Computation	20
7. CONCLUSION	22
BIBLIOGRAPHY	23

ACKNOWLEDGEMENT

We express our deep sense of gratitude to **Dr. R.Rudramoorthy**, Principal of PSG College of Technology, for having provided the necessary environments to carry out our project successfully.

We profusely thank, **Dr. R.Venkatesan**, Head of the Department, Computer Science and Engineering and Program Coordinator **Dr. G.R.Karpagam** who has greatly helped in the success of the project, by providing us with the necessary facilities required.

We express our gratitude to **Mrs.C.Kavitha** our guide, Assistant Professor(Sr.Gr.), Computer Science and Engineering, PSG College of Technology, for admitting us to proceed the project very effectively.

We extend our thanks to all our department staff, our friends and Librarian for their timely help and support to complete this project report successfully.

SYNOPSIS

Nowadays providing effective business analytics tools and technologies to the enterprise is a top priority for CEOs for good reason. Effective business analytics from basic reporting to advanced data mining and predictive analysis allows analysts and business users to extract insights from corporate data that then translated into action, deliver high levels of efficiency and profit to the enterprise.

The financial statement fraud involves senior management who are in the unique position to perpetrate fraud by overriding controls and acting in conclusion with other employees. When fraud occurs at lower levels in an organization, individuals may not initially realize that they are committing fraud.

Hence the project aims to take a radically different approach to data processing and analytics , by combining with sophisticated analytics to predict fraudulent companies. This facilitates customers to identify the status of companies from the annual reports published by them.

Annual reports of the companies are parsed and a score card is prepared. Based on this the status of the company can be identified.

LIST OF FIGURES

FIG NO.	NAME	PAGE NO.
4.1	Activity Diagram	9
4.2	Use case Diagram	10
5.1	Proposed system Design	11
6.1	Score card Preparation	17
6.1.1	Score card	18
6.2	Feature Selection	18
6.2	Classification Result	19

LIST OF TABLES

TABLE NO	NAME	PAGE NO
5.4	Result from Natural Extractor	13
6.1	Summary of Dataset	19
6.2	AccuracyComputation	20

CHAPTER 1

INTRODUCTION

The false financial statements are increasing frequently over the last few years. Falsifying financial statements primarily consists of manipulating elements by overstating assets, sales and profit, or understating liabilities, expenses, or losses. The Management fraud therefore, can be defined as deliberate fraud committed by management that injures investors and creditors through misleading financial statements.

The financial statement audit is a monitoring mechanism that helps reduce information asymmetry and protect the interests of the principals, specifically, stockholders and potential stockholders, by providing reasonable assurance that management's financial statements are free from material misstatements. Predictive analysis allows analysts and business users to extract insights from corporate data that then translated into action, deliver high levels of efficiency and profit to the enterprise.

Detecting management fraud is a difficult task when using normal audit procedures. The three underlying reasons are shortage of knowledge regarding characteristics of management fraud, lack of auditor expertise to detect fraud due to its infrequency and deception of auditors by the managers. These limitations suggest that there is a need for additional analytical procedures for the effective detection of management fraud

Hence the project aims to take a radically different approach to data processing and analytics, by combining with sophisticated analytics to predict fraudulent companies. Annual reports of the companies are parsed and a score card is prepared based on semantics. Features extracted from the score card can be used to study performance of the companies. This facilitates customers to identify the status of companies from the annual reports published by them.

1.1 MOTIVATION

To increase the foreign/corporate investments, it is necessary to improve the trust factor of companies from investor's perspective.

1.2 OBJECTIVE

The key objective is to find the financial fraudulent data provided by the companies and to provide an assurance to principals that, management's financial statements are free from material misstatements.

1.3 NEED FOR THE PROJECT

Apart from serving as an effective business analytical tool which extracts insight of corporate data, this project also delivers higher level of efficiency and profit.

1.4 ADVANTAGE

This project offers a wide range of benefits to customers by identifying fraudulent statements, providing an easier way to credit a firm and thereby increase their profitability. It also demonstrates an improved agility in analysing huge sets of data.

CHAPTER 2

LITERATURE SURVEY

2.1 PREDICTING FRAUDULENT FINANCIAL STATEMENTS WITH MACHINE LEARNING TECHNIQUES

Sotiris Kotsiantis, Euaggelos Koumanakos, Dimitris Tzelepis and Vasilis Tampakas[1]

The aim of this study is to investigate the usefulness and compare the performance of machine learning techniques in detecting fraudulent financial statements by using published financial data. A representative algorithm for each learning technique was used. The K2 algorithm was used to represent Bayesian networks. Ripper was used for rule learners and SMO algorithm was used in the representative of the SVMs. The results indicate that published financial statement data contains falsification indicators. The effectiveness of machine learning techniques in detecting firms that issue fraudulent financial statements (FFS) and deals with the identification of factors associated to FFS.

2.2 DETECTING FINANCIAL STATEMENT FRAUD

Johan L. Perols [2]

The goal of this dissertation is to improve financial statement fraud detection using a cross functional research approach. The efficacy of financial statement fraud detection depends on the classification algorithms and the fraud predictors used and how they are combined.

Essay I introduces IMF, a novel combiner method classification algorithm. The results show that IMF performs well relative to existing combiner methods over a wide range of domains. Essay II develops three novel fraud predictors: total discretionary accruals, meeting or beating analyst forecasts and unexpected employee productivity. The results show that the three variables are significant predictors of fraud. Hence Essay II provides insights into conditions under which fraud is more likely to occur, incentives for fraud and how fraud is committed and can be detected. Essay III compares the utility of artifacts developed in the broader research streams to which the first two essays contribute classification algorithm and fraud predictor research in detecting financial statement fraud.

2.3 PREVENTION AND DETECTION OF FINANCIAL STATEMENT FRAUD-AN IMPLEMENTATION OF DATA MINING FRAMEWORK

Rajan Gupta, Nasib Singh Gill [3]

Prevention and detection of financial statement fraud has become a major concern for almost all organisations globally. The informative variables have been used for implementing association rule mining for prevention and three predictive mining techniques namely Decision Tree, Naïve Bayesian Classifier, Genetic programming for detection of financial statement fraud. Rule Engine module of the framework generated seven association rules.

2.4 DETECTION OF FRAUDULENT FINANCIAL STATEMENTS THROUGH THE USE OF DATA MINING TECHNIQUES

Charalambos T. Spathis[4]

Three alternative models were built, each based on a different method. The first model includes decision tree model which was constructed using the Sipina Research Edition software. In second level splitters, two variables associated with profitability (NPTA and EBIT) were used. No fraud companies with a high z score present high profitability,Whereas fraud companies with a low z score present low profitability. In the third experiment we developed a Bayesian Belief Network (BBN). The software we used was the BN Power Predictor. This software is capable of learning a classifier from data.

2.5 MULTI-INSTANCE LEARNING FOR PREDICTING FRAUDULENT FINANCIAL STATEMENTS

Sotiris Kotsiantis[5]

Multi instance learning technique can facilitate auditors in accomplishing in the task of management fraud detection.This paper uses supervised machine learning methodology where each object in the set of training examples arelabelled and the problem is to learn a hypothesis that can accuratelypredicts the labels of the unseen objects. The training set could represent the data for companies which are known to have survived or gone bankrupt.

2.6 A REVIEW OF FINANCIAL ACCOUNTING FRAUD DETECTION BASED ON DATA MINING TECHNIQUES

AnujSharma,Prabin Kumar Panigrahi[6]

The data mining algorithms including statistical test, regression analysis, Neural Network, decision tree and Bayesian network for financial accounting fraud detection. Regression Analysis is widely used for fraud detection since it has great explanation ability. Different regression model used by researchers are Logit, Step-wise Logistic, UTADIS and EGB2 etc. Neural Networks are important tool for data mining. The advantages of Neural Network are that there are no strict requests for data and it has a strong generalization and adjustment. After correct allocation and proper training, Neural Network may perform greatclassification comparing with regression model. But due to special inner hidden structure, it is impossible to track the formation process of output conclusion. There are other issues also related with Neural Network like no clear explanation on connecting weight, complex accuracy and statistical reliability checking procedure, and lack of explanation.

This paper suggests that using only financial statements data may not be sufficient for detections of fraud. The importance of data mining techniques in the detection of financial fraud has been recognized.

CHAPTER 3

REQUIREMENTS

3.1 REQUIREMENT ANALYSIS

Hardware specifications

Processor	:	Intel Pentium dual core 2.40 GHz
Hard disk	:	500 GB
Memory	:	4 GB RAM
Keyboard	:	Standard

Software specifications

Operating System	:	Microsoft windows 7
Languages	:	Java
IDE	:	NetBeans 7.1.2
Libraries	:	ApacheTikka

3.2 FEASIBILITY ANALYSIS

This is done to assess the feasibility of the project. It is used to determine if the software to be built will meet the scope and requirements of the project.

3.2.1 Technical Feasibility

Technical feasibility analysis addresses the issues of technology used, portability and performance considerations of the technology chosen and its defects. The technologies used here are state of the art technologies. Use of Java gives robustness and support for portability.

3.2.2 Financial Feasibility

These analyses the cost factor involved with the project. This project does not involve a considerable cost factor as the supporting software (Jdk, Netbeans, Matlab) used are open source and free to download.

3.2.3 Schedule Feasibility

This addresses the main concern whether the project can be completed on time. A detailed time plan was prepared at the time of analysis and by following the schedule of the project, it was completed within time.

3.2.4 Resource Feasibility

This addresses the issue of resources required to implement the project. As far as this project is concerned there was no problem with the availability and setting up of needed resources.

CHAPTER 4

SYSTEM DESIGN

Systems design is the process of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to product development. There is some overlap with the disciplines of systems analysis, systems architecture and systems engineering. If the broader topic of product development blends the perspective of marketing, design, and manufacturing into a single approach to product development, then design is the act of taking the marketing information and creating the design of the product to be manufactured. Systems design is therefore the process of defining and developing systems to satisfy specified requirements of the user. Figure 4.1 and 4.2 demonstrates the system architecture in detail. The main steps of the system are the following:

4.1 ACTIVITY DIAGRAM

Activity diagram is basically a flow chart to represent the flow from one activity to another activity and captures the dynamic behaviour of the system. The activity can be described as an operation of the system. Activity diagrams deal with all type of flow control by using different elements like fork, join etc. Boxes of the activity diagram indicate the tasks and the arrows show the relationships.

In this project, an activity-on-node diagram has been designed after identifying the main tasks in the process. It starts with the collection of annual reports published by IT firms, which are parsed and framed into a dictionary in parallel which is indicated by the fork and join elements. The next task in the process is Scorecard preparation after which the classifier is trained and new inputs in the form of reports are run on the classifier to obtain the final result.

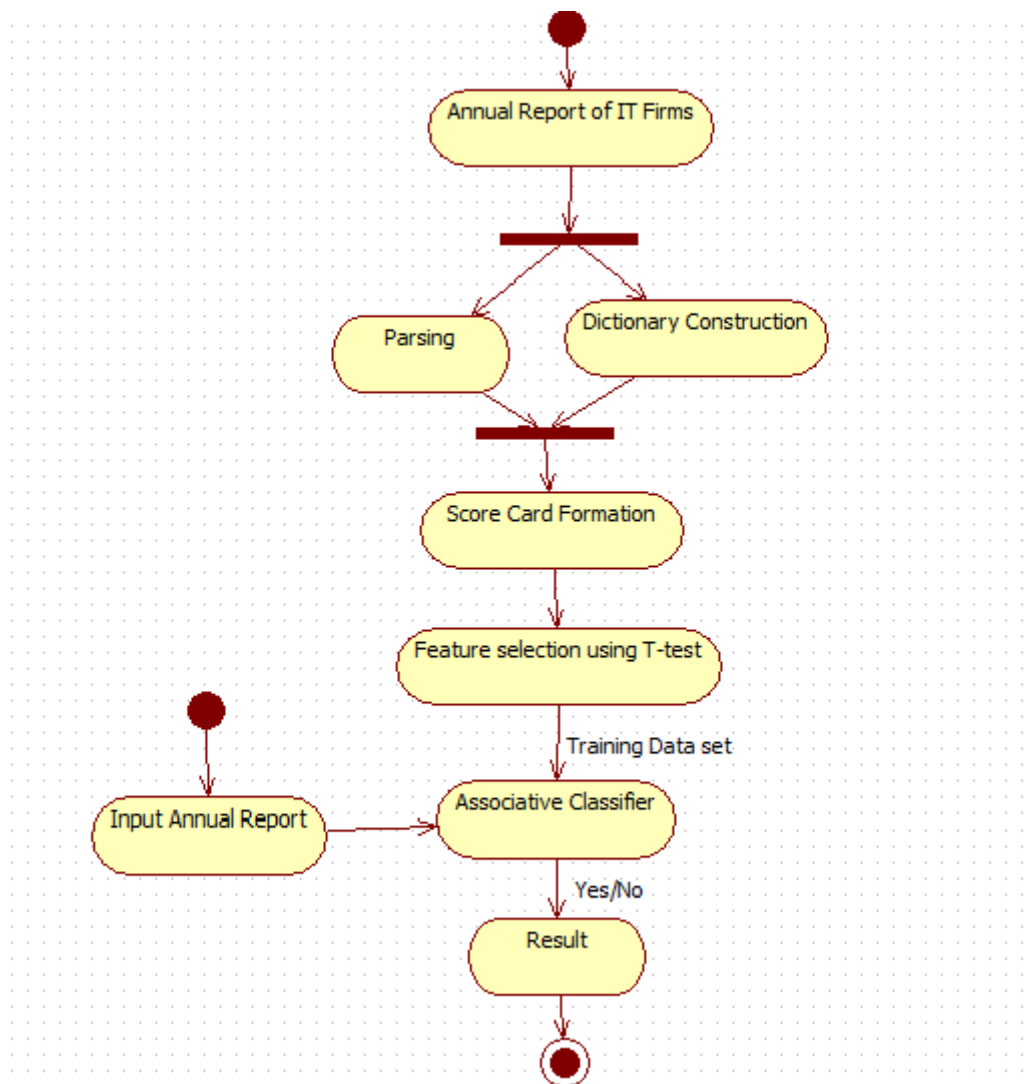


Figure 4.1 Activity Diagram

4.2 USE CASE DIAGRAM

A use case diagram is used to model the system/subsystem of an application and captures a particular functionality of a system. In its simplest definition, it is a representation of a user's interaction with the system and depicting the specifications of a use case. The internal and external agents that are responsible for making the interactions are known as actors. So use case diagrams consist of actors, use cases and their relationships.

The system under consideration in our project is Fraudulent Financial Detection system upon which a human user behaves as the actor. The system comprises of the five main use cases that have been identified and placed within oval shapes, and their relationship with the user is marked using straight lines. Collection of Annual Report use

case is included in the parsing use case because to perform parsing, the reports need to be gathered. Similarly, Dictionary construction is included in Scorecard preparation.

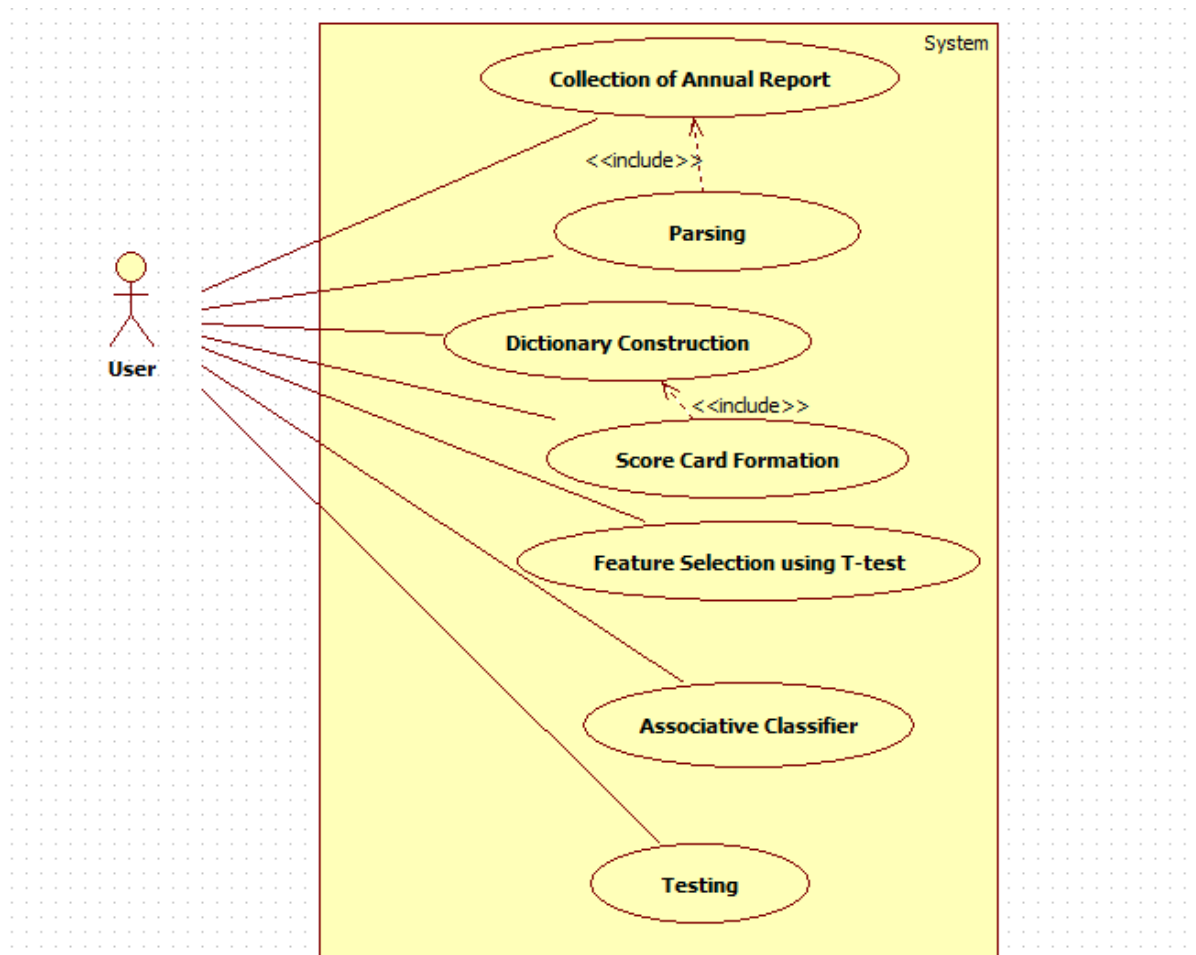


Figure 4.2 Use Case Diagram

CHAPTER 5

SYSTEM IMPLEMENTATION

5.1. REPORT PROCESSING

The annual reports for the IT firms which is available in PDF formats are collected and it is parsed. Dictionary is constructed based on the technique called Concept Extraction. The basic concepts behind each questionnaire provided by the government are extracted. The dictionary is then processed with annual reports to prepare the score card.

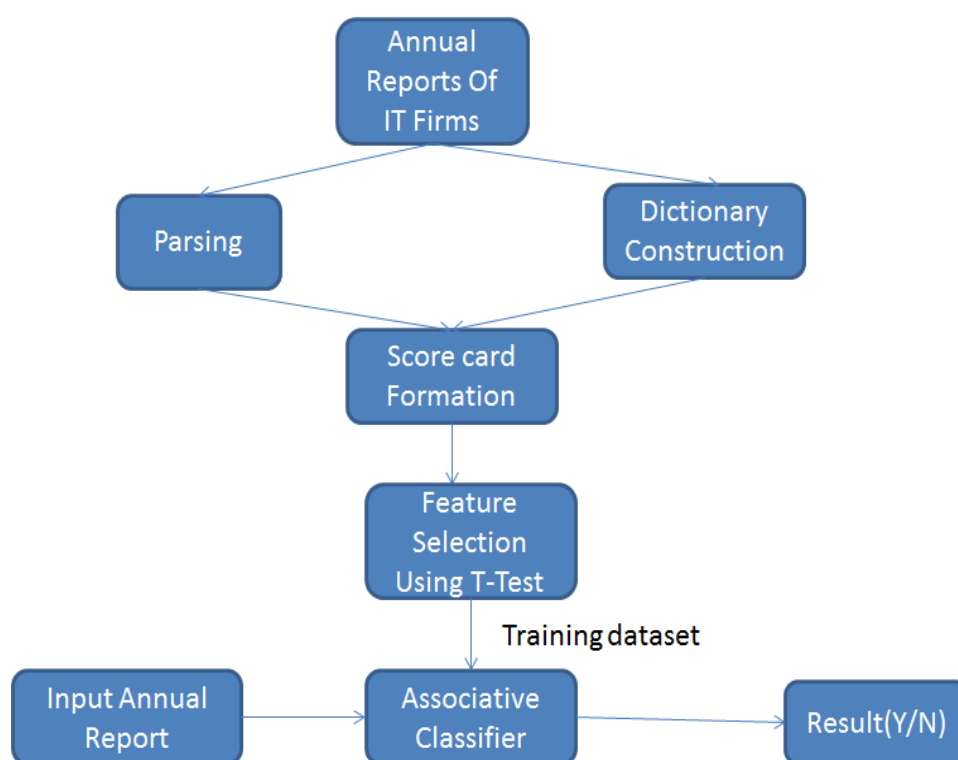


Figure 5.1 Proposed System

5.2 PARAMETERS

The government questionnaire has the following parameters. In each parameter there are set of questions which is used in the preparation of dictionary and to update score card.

- Governance Parameters.
- Board matters
- Nomination matters
- Remuneration matters
- Audit matters
- Communication

5.3 PARSING

The Apache Tika toolkit detects and extracts metadata and structured text content from various documents using existing parser libraries. It hides the complexity of different file formats and parsing libraries while providing a simple and powerful mechanism for client applications to extract structured text content and metadata from all sorts of documents. The **org.apache.tika.parser.Parser** interface is the key concept of Apache Tika. The first argument is an `InputStream` for reading the document to be parsed. If this document stream can not be read, then parsing stops and the `IOException` is passed up to the client application. If the stream can be read but not parsed then the parser throws a `TikaException`.

The final argument to the `parse` method is used to inject context-specific information to the parsing process. This is useful for example when dealing with locale-specific date and number formats in Microsoft Excel spreadsheets. Another important use of the parse context is passing in the delegate parser instance to be used by two-phase parsers like the `PackageParser` subclasses. Some parser classes allow customization of the parsing process through strategy objects in the parse context.

5.4 DICTIONARY CONSTRUCTION

The dictionary is constructed based on the “Concept Extraction” technique. It is the most powerful and flexible technology for detecting key ideas and trends. Syntactic analysis is used to identify not only noun phrases, but also verb phrases, adjectival phrases and more, according to user’s needs.

Concepts are minimum fragments of text which refer to **objects** or **ideas**. The service can be used to extract objects and ideas from any type of text, either high-quality ones (news, legislation...) or colloquial ones (forums, blogs, chats, social media...).

NaturalExtractor uses Deep Linguistic Analysis based on grammars, which allows for a rich variety of concept detection adapted according to user’s needs:

- Simple concepts: “checkingaccount”, “Corporate policy”
- Compound or nested concepts: “Company’s checking account”
- Combinations of the above: “account”, “checking account”, “checking account of the bank”, “bank”.

The linguistic analysis also makes it possible to extract concepts based on syntactic phrases:

- Noun phrases: “Are criteria used for **individual director** performance evaluation disclosed?”
- Adjectival phrases: “Are **criteria** used for individual director performance evaluation **disclosed**?”
- Verb phrases: “**Are** criteria used for individual director performance evaluation **disclosed**?”

The results from NaturalExtractor for few examples of such questions are,

Table 5.4 Result from Natural Extractor

QUESTIONNAIRE	KEYWORDS
Are the type of material transactions that must be approved by the board disclosed?	Materials transaction approval
Was appraisal of board performance conducted?	Appraisal of board performance details

Statement of Company's philosophy on code of governance.	Corporate governance philosophy
Do the criteria used for evaluating board performance include financial measures such as the company's share price performance or total shareholder return?	Criteria for board evaluation, financial measures
Are criteria used for individual director performance evaluation disclosed?	Criteria for individual director evaluation

5.5 SCORE CARD PREPARATION

The score card contains 115 features in which each one has an assigned score. The score of the features falls under three categories, each has different scores such as 1,2 and 5. It takes the annual report and dictionary as input in order to perform matching.

N-gram matching technique is used for the comparison of features with annual reports. When the match is found, the corresponding score is assigned to that particular feature. Otherwise the feature is assigned with null value.

N-gram models are widely used in statistical natural language processing. For parsing, words are modelled such that each n-gram is composed of n words. For sequences of words, the trigrams that can be generated from "Statement of Company's philosophy on code of governance." are "# the company", "statement of company", "company philosophy #", "code of governance", "philosophy on #" and "statement of #". Initial step is to pre-process the strings to remove spaces, most simply collapsewhitespace to a single space while preserving paragraph marks. Punctuation is also commonly reduced or removed by pre-processing

5.6 FEATURE SELECTION

5.6.1 T-Test Evaluation

A t-test is used to compare whether the two groups have different average values. It can be done by using the MATLAB built-in function called `mattest()`. The general form of the function is given below.

```
PValues = mattest(DataX, DataY)
[PValues, TScores] = mattest(DataX, DataY)
```

DataMatrix object is a matrix of feature values where each row corresponds to a feature and each column corresponds to a company. DataX and DataY must have the same number of rows and are assumed to be normally distributed in each class with equal variances.

Where DataX contains data from well growing companies and DataY contains data from a low performing companies. For example, DataX could be feature values from TCS, Infosys, and DataY could be feature values from Zenith Birla and KingFisher Airlines.

The output from the above function is a matrix which contains a column vector of p-values and a column vector of t-scores. Then the p-values column vector is arranged in descending order in order to find the average similarity between the features. Now the top p-values which satisfies the threshold are taken as a new feature set. So that the common features are extracted from the whole 115 features.

5.7 ASSOCIATIVE CLASSIFICATION

Association classification is a recent and rewarding technique that combines the methodology of association and classification. This generally involves two stages:

1. Generate class association rules from a training data set.
2. Classify the test data set into predefined class labels.

In short, Association classification utilizes the association rule discovery techniques to construct classification systems.

An Association Rule is a rule of the form

$$\textit{cot and bed} \rightarrow \textit{pillow}$$

where 'cot and bed' is called the rule body and pillow the head of the rule. It associates the rule body with its head. The following example expresses the fact that people who buy cot and bed are likely to buy pillow too.

An association rule can be defined as: Let D be a database consisting of one table over n attributes $\{a_1, a_2, \dots, a_n\}$. Let this table contain k instances. Let d be a database record. d satisfies an item set $X \subseteq \{a_1, a_2, \dots, a_n\}$ if $X \subseteq d$. An association rule is an implication $X \rightarrow Y$ where $X, Y \subseteq \{a_1, a_2, \dots, a_n\}$. The use of association rules for classification is restricted to problems where the instances can only belong to a discrete number of classes. The reason is that association rule mining is only possible for nominal attributes. However, association rules in their general form cannot be used directly. Their definition has to be restricted. The head Y of an arbitrary association rule $X \rightarrow Y$ is a disjunction of items. Every item which is not present in the rule body may occur in the head of the rule. The rules that are capable of assigning a class membership can be used as rules for classification. Therefore the head Y of a class association rule $X \rightarrow Y$ is restricted to one item. The attribute of this attribute-value-pair has to be the class attribute.

According to this, a class association rule is of the form $X \rightarrow a_i$ where a_i is the class attribute and $X \subseteq \{a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n\}$

5.7.1 Steps in Associative Classification Rule Mining

The steps in associative classification can be summarized as follows:

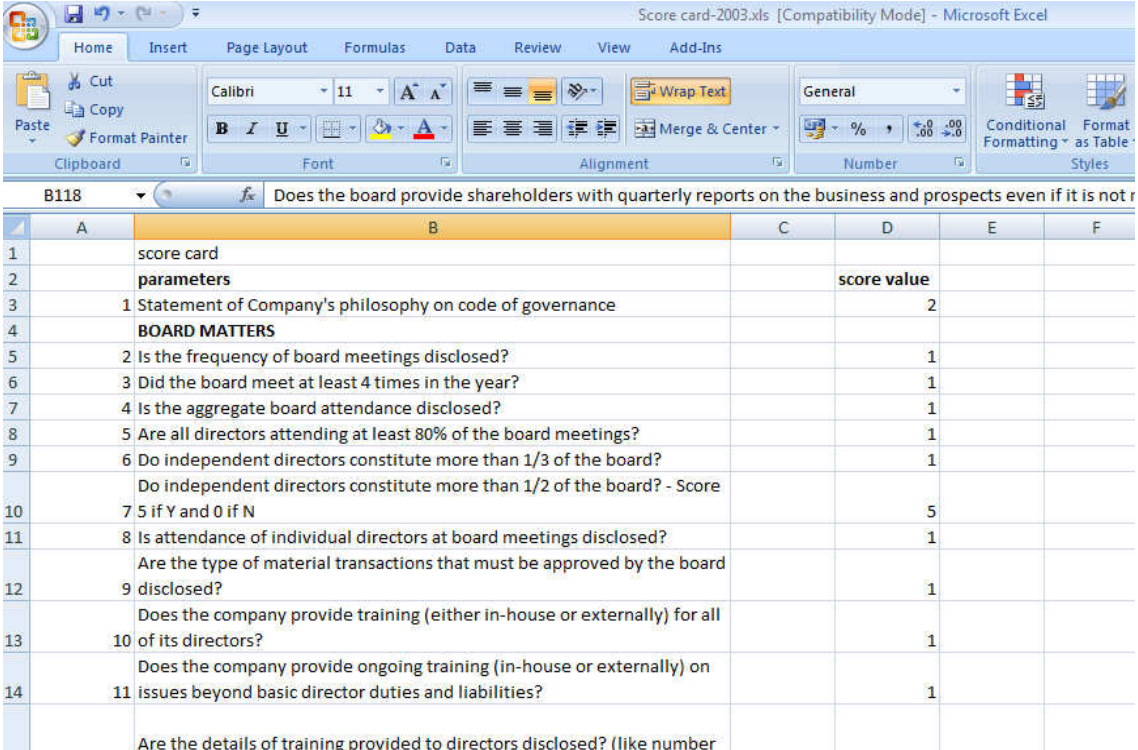
- The discovery of all frequent rule items from a training dataset which yields class association rules.
- Ranking of the Class Association Rules (CAR) and pruning them based on the measures of interestingness (Support, Confidence) and their threshold values.
- The selection of one subset of CARs to form the classifier.
- Measuring the quality of the derived classifier on test data objects.

CHAPTER 6

EXPERIMENTAL RESULTS

6.1 SCORE CARD FORMATION

The sample score card with 115 features and their assigned score is given in Figure 6.1.



The screenshot shows a Microsoft Excel spreadsheet titled "Score card-2003.xls [Compatibility Mode] - Microsoft Excel". The spreadsheet contains a table with 6 columns (A-F) and 15 rows. The table is a score card with 115 features and their assigned scores. The features are listed in column A, and the scores are listed in column D. The table is titled "score card" and "parameters" in row 1. The features are listed in rows 2-15, and the scores are listed in rows 2-15. The table is titled "score card" and "parameters" in row 1. The features are listed in rows 2-15, and the scores are listed in rows 2-15.

	A	B	C	D	E	F
1	score card					
2	parameters			score value		
3	1 Statement of Company's philosophy on code of governance			2		
4	BOARD MATTERS					
5	2 Is the frequency of board meetings disclosed?			1		
6	3 Did the board meet at least 4 times in the year?			1		
7	4 Is the aggregate board attendance disclosed?			1		
8	5 Are all directors attending at least 80% of the board meetings?			1		
9	6 Do independent directors constitute more than 1/3 of the board?			1		
10	7 5 if Y and 0 if N			5		
11	8 Is attendance of individual directors at board meetings disclosed?			1		
12	9 Are the type of material transactions that must be approved by the board disclosed?			1		
13	10 Does the company provide training (either in-house or externally) for all of its directors?			1		
14	11 Does the company provide ongoing training (in-house or externally) on issues beyond basic director duties and liabilities?			1		
	Are the details of training provided to directors disclosed? (like number					

Figure 6.1 Score Card Representation

The score card generated for companies such as CTS, WIPRO, TCS, HCL, KingFisher Airlines, ITC and Zenith Brila is given in Figure 6.1.1

Book1.xlsx - Microsoft Excel

HomeInsertPage LayoutFormulasDataReviewViewAdd-Ins

Paste

Clipboard...

Courier New10.5A⁺A⁻

B

I

U

Font

Alignment

General

Number

Conditional Formatting

Format as Table

Cell Styles

Cells

Σ

Editing

Sort & Filter

Find & Select

S14

fx

1

	E	F	G	H	I	J	K	L	M	N	O	P	Q	R		
1	2007-08				2008-09				2009-10				2010-11			
2	WIPRO	TCS	HCL	CTS	WIPRO	TCS	HCL	CTS	WIPRO	TCS	HCL	CTS	WIPRO	TCS	HCL	
3	2	2	2	2	2	2	2	2	2	2	2	2	2	2		
4																
5	1	1	1	0	1	1	1	0	1	1	1	0	1	1		
6	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
7	1	1	1	0	1	1	1	0	1	1	1	0	1	1		
8	0	1	1	0	0	1	1	0	1	1	1	0	0	1		
9	5	5	5	5	5	5	5	5	5	5	5	5	5	5		
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
11	0	1	1	0	0	1	1	0	0	1	1	0	0	1		
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
13	1	1	1	0	1	0	1	0	1	0	1	0	0	1		

Ready

Sheet1Sheet2Sheet3

100%

Figure 6.1.1 Score Card

6.2 Feature Selection

Among 115 features, 76 features are selected based on T-test evaluation and are listed in Figure 6.2

```

board meeting held atleast 4 times a year/number of meetings / frequency of mee
independent directors
induction program/orientation program
directors on the board/number of directorships and committee chairmanships in o
names and categories of the directors on the board/name of the director/ catego
composition of the board/conformity with clause 49 of the listing agreements/bo
chairman/director of board/non-independent/non-executive
chairman*s responsibility/board meetings/ administrative matters/encourage rela
managing director/nomination/re-election/regular intervals/every 3 years
services of managing directors and executive directors/six months* notice/notic
independent directors/separate access to secretary/independent access
write to the company secretary/role of secretary/good information flow/director
nominations committee of directors/independent directors/chairman of nominating
time to time/atleast annually/atleast yearly/ nominations committee/nominating
affairs/position development/adequacy of time/review/r/time spent by director
disclosures/attendance general board members meeting/nominating committee membe
remuneration policy/board evaluation/conducted by external party/ conducted by
economic value added analysis/remuneration policy/share price/criteria for boar
director evaluation /individual performance of board members
director evaluation process/director performance evaluation
economic value added analysis/remuneration policy/share price/criteria for boar
list of / composition of remuneration committee
independent directors/ remuneration committee
terms of reference of remuneration committee recommend/approve
attended/held/attendance of remuneration committee members
out of pocket expense/specific remuneration packages/remuneration packages
executive directors/commission/performance of company and each executive direct
executive directors/ performance/commission
services/terminated/six months* notice/ company paying six months* salary
annual appraisal process/benefits and economic value added analysis/retirement
details of the remuneration/ benefits/perquisites/allowances/commission
details of the remuneration/ benefits/perquisites/allowances/commission

```

Figure 6.2 Feature Selection

6.3 Associative Classification

By using the above selected features, training data is given and then the association rules are generated. Now the testing is done from those rules.

[illegible]

Figure 6.3 Classification Result

6.4 Summary Of Dataset

The Annual reports from the firms such as TCS, Wipro, Zenith Birla and KingFisherAirlines,etc., are taken. The details on the count of documents are given in Table 6.1

Table 6.1 Summary Of Dataset

Name Of The Company	Number of Documents
CTS	7
TCS	7

Name Of The Company	Number of Documents
Wipro	5
HCL	7
Zenith Birla	6
KingFisherAirLines	5
M-Phasis	6
ITC	7
Total Number Of Documents	50

The first column represents the name of the company. The second column gives the number of documents in each company. The last row represents the total number of documents.

6.5 ACCURACY COMPUTATION

Accuracy measures the ability of the classifier to correctly classify unlabeled data. It is the ratio of the number of correctly classified data over the total number of given transactions in the test dataset.

$$\text{Accuracy} = \frac{\text{Number of correctly predicted test data}}{\text{total number of test data}}$$

Table 6.2 Accuracy computation

	Accuracy(in percentage)	Number Of Rules
Without Feature Selection	75.0	16
With Feature Selection	87.5	16

The second row gives the accuracy calculated for 115 features. The third row gives the accuracy calculated for 76 features(which is derived from 115 features using T-test).

From the above tabulated data, accuracy computed before feature selection is lesser than the accuracy with feature selection. Thus feature selection by T-test gives enhanced results.

CHAPTER 7

CONCLUSION

Every day, news of financial statement fraud is adversely affecting the economy worldwide. Prediction of financial statement fraud would be of great value to the organizations throughout the world. Considering the need of such a mechanism, a data mining framework is to be employed for prediction of financial statement fraud.

Thus our project aims to expand the understanding of how deceivers use language differently than truth tellers, particularly in high-stakes, real-world environments such as financial markets. It takes a radically different approach to data processing and analytics, by combining with sophisticated analytics to predict fraudulent companies. This facilitates customers to identify the status of companies from the annual reports published by them.

BIBLIOGRAPHY

- [1]S. Kotsiantis,Euaggelos koumanakos,Dimitris Tzelepis and Vasilis Tampakas.,“Predicting fraudulent financial statements with machine learning techniques”, International Journal of Computational Intelligence, Vol. 3, No. 2, pp.104-110, 2006.
- [2]Johan L. Perols.,“Detecting financial statement fraud”,International Journal of Computer Applications,2008.
- [3]Rajan Gupta, Nasib Singh Gill., “Prevention and Detection of Financial Statement fraud”, International Journal of Advanced Computer Science and Applications, Vol. 3, No. 8, 2012.
- [4]Efsthios Kirkos, Charalambos Spathis, Yannis Manolopoulos.,“Detection Of Fraudulent Financial Statements Through The Use Of Data Mining Techniques”,2nd International Conference on Enterprise Systems and Accounting,2005.
- [5]Sotiris Kotsiantis,“Multi-Instance Learning For Predicting Fraudulent Financial Statements”,Third International Conference In Convergence and Hybrid Information Technology,, Vol. 1,2008.
- [6]Anuj Sharma,Prabin Kumar Panigrahi.,“A Review Of Financial Accounting Fraud Detection Based On Data Mining Technique”, International Journal of Computer Applications (0975 – 8887)Volume 39– No.1, 2013.
- [7]O. Persons, "Using financial statement data to identify factors associated with fraudulent financing reporting," Journal of Applied Business Research,vol. 11, pp. 38-46, 1995.
- [8]C. Spathis, "Detecting false financial statements usingpublished, data: some evidence from Greece,"Managerial Auditing Journal, vol. 17, no. 4, pp.179-191,2002.
- [9]J. W. Lin, M. I. Hwang, and J. D. Becker, "A fuzzyneural network for assessing the risk of fraudulentfinancial reporting," Managerial Auditing Journal,Vol. 18, pp. 657-665,2003.

CONTENTS

CHAPTER	Page No.
Acknowledgement.....	(i)
Synopsis.....	(ii)
List of figures.....	(iii)
List of tables.....	(iv)
1. INTRODUCTION.....	1
1.1. Motivation	1
1.2. Objective	2
1.3. Need for the project	2
1.4. Advantage	2
2. LITERATURE SURVEY.....	3
2.1 Predicting fraudulent financial statements with machine learning techniques	3
2.2 Detecting financial statement fraud	3
2.3. Prevention and detection of financial statement fraud-an implementation of data mining framework	4
2.4 Detection of fraudulent financial statements through the use of data mining techniques	4
2.5 Multi-instance learning for predicting fraudulent financial statement	4
2.6 A review of financial accounting fraud detection based on data mining Techniques	5
3. REQUIREMENTS.....	6
3.1. Requirement Analysis	6
3.2. Feasibility Analysis	7
3.2.1 Technical Feasibility	7
3.2.2 Financial Feasibility	7
3.2.3 Schedule Feasibility	7
3.2.4 Resource Feasibility	7

4. SYSTEM DESIGN.....	8
4.1. Activity Diagram	8
4.2. Usecase Diagram	9
5. SYSTEM IMPLEMENTATION.....	11
5.1. Report Processing	11
5.2. Parameters	12
5.3. Parsing	12
5.4. Dictionary Construction	13
5.5. Score card Preparation	14
5.6. Feature Selection	15
5.6.1 T-Test Evaluation	15
5.7. Associative Classification	15
5.7.1 Steps in Associative Classification Rule Mining	16
6. EXPERIMENTAL RESULTS.....	17
6.1. Score card formation	17
6.2. Feature Selection	18
6.3. Associative Classification	19
6.4. Summary of Dataset	19
6.5. Accuracy of Computation	20
7. CONCLUSION	22
BIBLIOGRAPHY	23

ACKNOWLEDGEMENT

We express our deep sense of gratitude to **Dr. R.Rudramoorthy**, Principal of PSG College of Technology, for having provided the necessary environments to carry out our project successfully.

We profusely thank, **Dr. R.Venkatesan**, Head of the Department, Computer Science and Engineering and Program Coordinator **Dr. G.R.Karpagam** who has greatly helped in the success of the project, by providing us with the necessary facilities required.

We express our gratitude to **Mrs.C.Kavitha** our guide, Assistant Professor(Sr.Gr.), Computer Science and Engineering, PSG College of Technology, for admitting us to proceed the project very effectively.

We extend our thanks to all our department staff, our friends and Librarian for their timely help and support to complete this project report successfully.

SYNOPSIS

Nowadays providing effective business analytics tools and technologies to the enterprise is a top priority for CEOs for good reason. Effective business analytics from basic reporting to advanced data mining and predictive analysis allows analysts and business users to extract insights from corporate data that then translated into action, deliver high levels of efficiency and profit to the enterprise.

The financial statement fraud involves senior management who are in the unique position to perpetrate fraud by overriding controls and acting in conclusion with other employees. When fraud occurs at lower levels in an organization, individuals may not initially realize that they are committing fraud.

Hence the project aims to take a radically different approach to data processing and analytics , by combining with sophisticated analytics to predict fraudulent companies. This facilitates customers to identify the status of companies from the annual reports published by them.

Annual reports of the companies are parsed and a score card is prepared. Based on this the status of the company can be identified.

LIST OF FIGURES

FIG NO.	NAME	PAGE NO.
4.1	Activity Diagram	9
4.2	Use case Diagram	10
5.1	Proposed system Design	11
6.1	Score card Preparation	17
6.1.1	Score card	18
6.2	Feature Selection	18
6.2	Classification Result	19

LIST OF TABLES

TABLE NO	NAME	PAGE NO
5.4	Result from Natural Extractor	13
6.1	Summary of Dataset	19
6.2	AccuracyComputation	20

CHAPTER 1

INTRODUCTION

The false financial statements are increasing frequently over the last few years. Falsifying financial statements primarily consists of manipulating elements by overstating assets, sales and profit, or understating liabilities, expenses, or losses. The Management fraud therefore, can be defined as deliberate fraud committed by management that injures investors and creditors through misleading financial statements.

The financial statement audit is a monitoring mechanism that helps reduce information asymmetry and protect the interests of the principals, specifically, stockholders and potential stockholders, by providing reasonable assurance that management's financial statements are free from material misstatements. Predictive analysis allows analysts and business users to extract insights from corporate data that then translated into action, deliver high levels of efficiency and profit to the enterprise.

Detecting management fraud is a difficult task when using normal audit procedures. The three underlying reasons are shortage of knowledge regarding characteristics of management fraud, lack of auditor expertise to detect fraud due to its infrequency and deception of auditors by the managers. These limitations suggest that there is a need for additional analytical procedures for the effective detection of management fraud

Hence the project aims to take a radically different approach to data processing and analytics, by combining with sophisticated analytics to predict fraudulent companies. Annual reports of the companies are parsed and a score card is prepared based on semantics. Features extracted from the score card can be used to study performance of the companies. This facilitates customers to identify the status of companies from the annual reports published by them.

1.1 MOTIVATION

To increase the foreign/corporate investments, it is necessary to improve the trust factor of companies from investor's perspective.

1.2 OBJECTIVE

The key objective is to find the financial fraudulent data provided by the companies and to provide an assurance to principals that, management's financial statements are free from material misstatements.

1.3 NEED FOR THE PROJECT

Apart from serving as an effective business analytical tool which extracts insight of corporate data, this project also delivers higher level of efficiency and profit.

1.4 ADVANTAGE

This project offers a wide range of benefits to customers by identifying fraudulent statements, providing an easier way to credit a firm and thereby increase their profitability. It also demonstrates an improved agility in analysing huge sets of data.

CHAPTER 2

LITERATURE SURVEY

2.1 PREDICTING FRAUDULENT FINANCIAL STATEMENTS WITH MACHINE LEARNING TECHNIQUES

Sotiris Kotsiantis, Euaggelos Koumanakos, Dimitris Tzelepis and Vasilis Tampakas[1]

The aim of this study is to investigate the usefulness and compare the performance of machine learning techniques in detecting fraudulent financial statements by using published financial data. A representative algorithm for each learning technique was used. The K2 algorithm was used to represent Bayesian networks. Ripper was used for rule learners and SMO algorithm was used in the representative of the SVMs. The results indicate that published financial statement data contains falsification indicators. The effectiveness of machine learning techniques in detecting firms that issue fraudulent financial statements (FFS) and deals with the identification of factors associated to FFS.

2.2 DETECTING FINANCIAL STATEMENT FRAUD

Johan L. Perols [2]

The goal of this dissertation is to improve financial statement fraud detection using a cross functional research approach. The efficacy of financial statement fraud detection depends on the classification algorithms and the fraud predictors used and how they are combined.

Essay I introduces IMF, a novel combiner method classification algorithm. The results show that IMF performs well relative to existing combiner methods over a wide range of domains. Essay II develops three novel fraud predictors: total discretionary accruals, meeting or beating analyst forecasts and unexpected employee productivity. The results show that the three variables are significant predictors of fraud. Hence Essay II provides insights into conditions under which fraud is more likely to occur, incentives for fraud and how fraud is committed and can be detected. Essay III compares the utility of artifacts developed in the broader research streams to which the first two essays contribute classification algorithm and fraud predictor research in detecting financial statement fraud.

2.3 PREVENTION AND DETECTION OF FINANCIAL STATEMENT FRAUD-AN IMPLEMENTATION OF DATA MINING FRAMEWORK

Rajan Gupta, Nasib Singh Gill [3]

Prevention and detection of financial statement fraud has become a major concern for almost all organisations globally. The informative variables have been used for implementing association rule mining for prevention and three predictive mining techniques namely Decision Tree, Naïve Bayesian Classifier, Genetic programming for detection of financial statement fraud. Rule Engine module of the framework generated seven association rules.

2.4 DETECTION OF FRAUDULENT FINANCIAL STATEMENTS THROUGH THE USE OF DATA MINING TECHNIQUES

Charalambos T. Spathis[4]

Three alternative models were built, each based on a different method. The first model includes decision tree model which was constructed using the Sipina Research Edition software. In second level splitters, two variables associated with profitability (NPTA and EBIT) were used. No fraud companies with a high z score present high profitability,Whereas fraud companies with a low z score present low profitability. In the third experiment we developed a Bayesian Belief Network (BBN). The software we used was the BN Power Predictor. This software is capable of learning a classifier from data.

2.5 MULTI-INSTANCE LEARNING FOR PREDICTING FRAUDULENT FINANCIAL STATEMENTS

Sotiris Kotsiantis[5]

Multi instance learning technique can facilitate auditors in accomplishing in the task of management fraud detection.This paper uses supervised machine learning methodology where each object in the set of training examples arelabelled and the problem is to learn a hypothesis that can accuratelypredicts the labels of the unseen objects. The training set could represent the data for companies which are known to have survived or gone bankrupt.

2.6 A REVIEW OF FINANCIAL ACCOUNTING FRAUD DETECTION BASED ON DATA MINING TECHNIQUES

AnujSharma,Prabin Kumar Panigrahi[6]

The data mining algorithms including statistical test, regression analysis, Neural Network, decision tree and Bayesian network for financial accounting fraud detection. Regression Analysis is widely used for fraud detection since it has great explanation ability. Different regression model used by researchers are Logit, Step-wise Logistic, UTADIS and EGB2 etc. Neural Networks are important tool for data mining. The advantages of Neural Network are that there are no strict requests for data and it has a strong generalization and adjustment. After correct allocation and proper training, Neural Network may perform greatclassification comparing with regression model. But due to special inner hidden structure, it is impossible to track the formation process of output conclusion. There are other issues also related with Neural Network like no clear explanation on connecting weight, complex accuracy and statistical reliability checking procedure, and lack of explanation.

This paper suggests that using only financial statements data may not be sufficient for detections of fraud. The importance of data mining techniques in the detection of financial fraud has been recognized.

CHAPTER 3

REQUIREMENTS

3.1 REQUIREMENT ANALYSIS

Hardware specifications

Processor	:	Intel Pentium dual core 2.40 GHz
Hard disk	:	500 GB
Memory	:	4 GB RAM
Keyboard	:	Standard

Software specifications

Operating System	:	Microsoft windows 7
Languages	:	Java
IDE	:	NetBeans 7.1.2
Libraries	:	ApacheTikka

3.2 FEASIBILITY ANALYSIS

This is done to assess the feasibility of the project. It is used to determine if the software to be built will meet the scope and requirements of the project.

3.2.1 Technical Feasibility

Technical feasibility analysis addresses the issues of technology used, portability and performance considerations of the technology chosen and its defects. The technologies used here are state of the art technologies. Use of Java gives robustness and support for portability.

3.2.2 Financial Feasibility

These analyses the cost factor involved with the project. This project does not involve a considerable cost factor as the supporting software (Jdk, Netbeans, Matlab) used are open source and free to download.

3.2.3 Schedule Feasibility

This addresses the main concern whether the project can be completed on time. A detailed time plan was prepared at the time of analysis and by following the schedule of the project, it was completed within time.

3.2.4 Resource Feasibility

This addresses the issue of resources required to implement the project. As far as this project is concerned there was no problem with the availability and setting up of needed resources.

CHAPTER 4

SYSTEM DESIGN

Systems design is the process of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to product development. There is some overlap with the disciplines of systems analysis, systems architecture and systems engineering. If the broader topic of product development blends the perspective of marketing, design, and manufacturing into a single approach to product development, then design is the act of taking the marketing information and creating the design of the product to be manufactured. Systems design is therefore the process of defining and developing systems to satisfy specified requirements of the user. Figure 4.1 and 4.2 demonstrates the system architecture in detail. The main steps of the system are the following:

4.1 ACTIVITY DIAGRAM

Activity diagram is basically a flow chart to represent the flow from one activity to another activity and captures the dynamic behaviour of the system. The activity can be described as an operation of the system. Activity diagrams deal with all type of flow control by using different elements like fork, join etc. Boxes of the activity diagram indicate the tasks and the arrows show the relationships.

In this project, an activity-on-node diagram has been designed after identifying the main tasks in the process. It starts with the collection of annual reports published by IT firms, which are parsed and framed into a dictionary in parallel which is indicated by the fork and join elements. The next task in the process is Scorecard preparation after which the classifier is trained and new inputs in the form of reports are run on the classifier to obtain the final result.

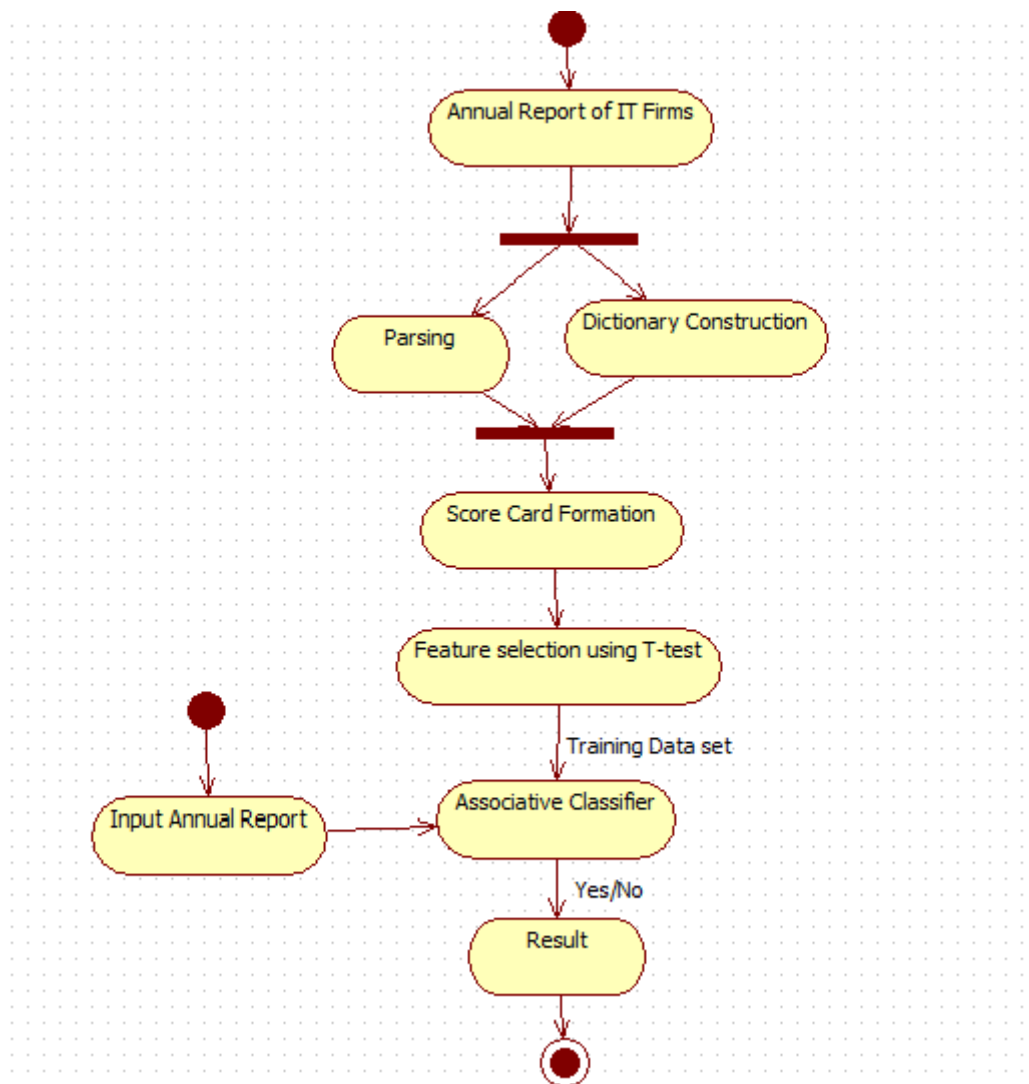


Figure 4.1 Activity Diagram

4.2 USE CASE DIAGRAM

A use case diagram is used to model the system/subsystem of an application and captures a particular functionality of a system. In its simplest definition, it is a representation of a user's interaction with the system and depicting the specifications of a use case. The internal and external agents that are responsible for making the interactions are known as actors. So use case diagrams consist of actors, use cases and their relationships.

The system under consideration in our project is Fraudulent Financial Detection system upon which a human user behaves as the actor. The system comprises of the five main use cases that have been identified and placed within oval shapes, and their relationship with the user is marked using straight lines. Collection of Annual Report use

case is included in the parsing use case because to perform parsing, the reports need to be gathered. Similarly, Dictionary construction is included in Scorecard preparation.

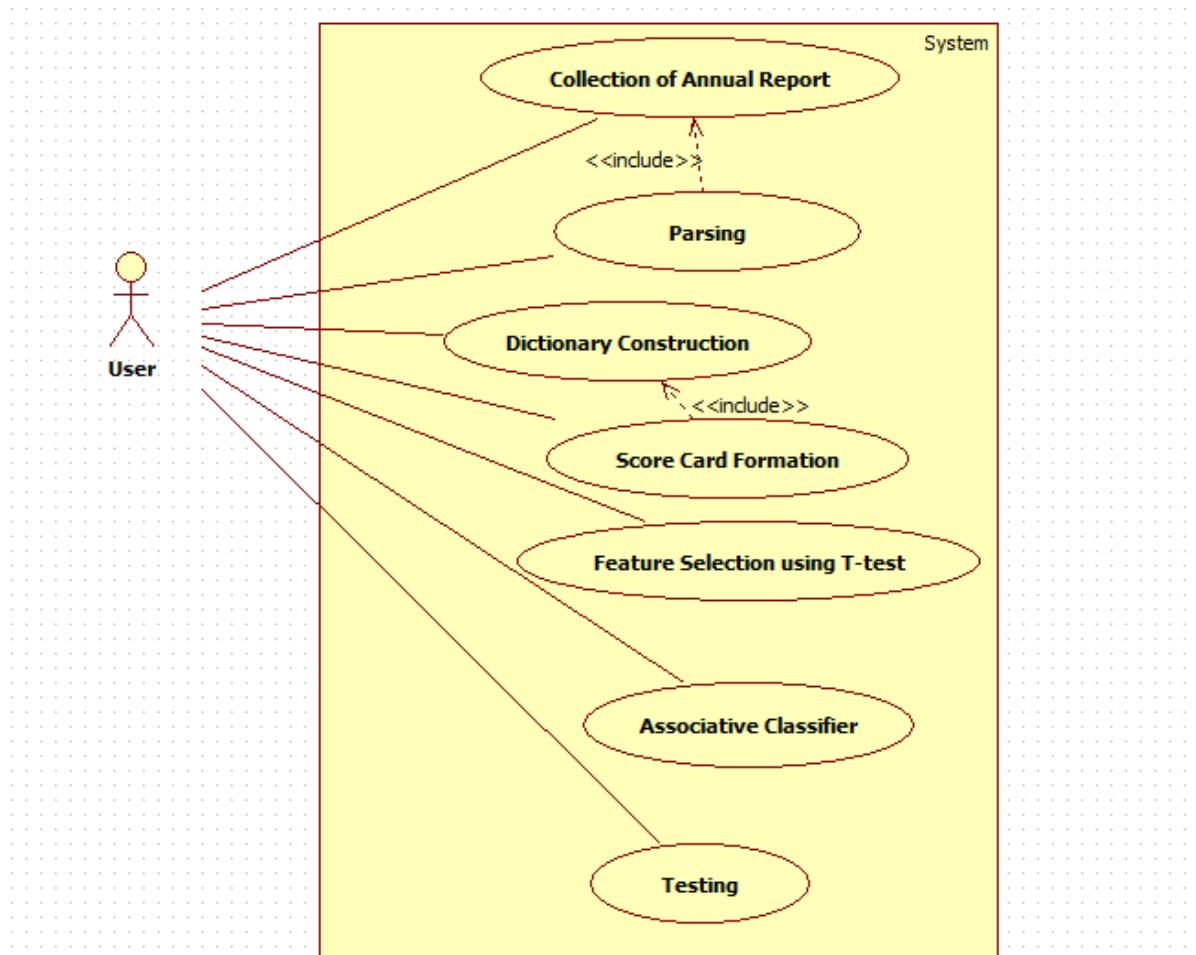


Figure 4.2 Use Case Diagram

CHAPTER 5

SYSTEM IMPLEMENTATION

5.1. REPORT PROCESSING

The annual reports for the IT firms which is available in PDF formats are collected and it is parsed. Dictionary is constructed based on the technique called Concept Extraction. The basic concepts behind each questionnaire provided by the government are extracted. The dictionary is then processed with annual reports to prepare the score card.

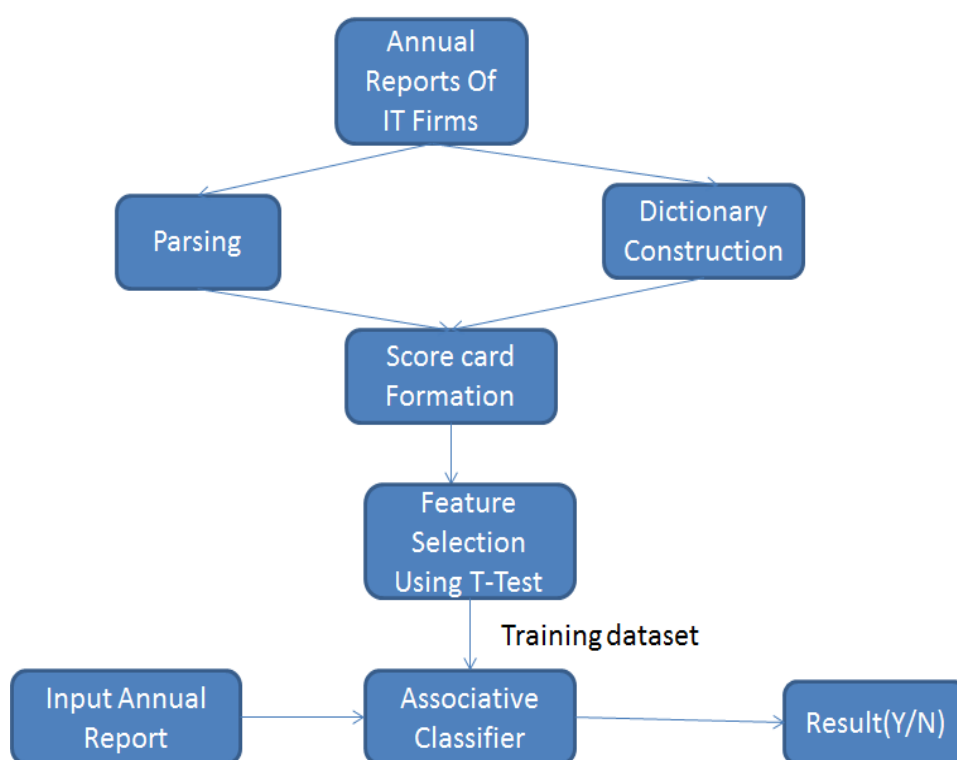


Figure 5.1 Proposed System

5.2 PARAMETERS

The government questionnaire has the following parameters. In each parameter there are set of questions which is used in the preparation of dictionary and to update score card.

- Governance Parameters.
- Board matters
- Nomination matters
- Remuneration matters
- Audit matters
- Communication

5.3 PARSING

The Apache Tika toolkit detects and extracts metadata and structured text content from various documents using existing parser libraries. It hides the complexity of different file formats and parsing libraries while providing a simple and powerful mechanism for client applications to extract structured text content and metadata from all sorts of documents. The **org.apache.tika.parser.Parser** interface is the key concept of Apache Tika. The first argument is an `InputStream` for reading the document to be parsed. If this document stream can not be read, then parsing stops and the `IOException` is passed up to the client application. If the stream can be read but not parsed then the parser throws a `TikaException`.

The final argument to the `parse` method is used to inject context-specific information to the parsing process. This is useful for example when dealing with locale-specific date and number formats in Microsoft Excel spreadsheets. Another important use of the parse context is passing in the delegate parser instance to be used by two-phase parsers like the `PackageParser` subclasses. Some parser classes allow customization of the parsing process through strategy objects in the parse context.

5.4 DICTIONARY CONSTRUCTION

The dictionary is constructed based on the “Concept Extraction” technique. It is the most powerful and flexible technology for detecting key ideas and trends. Syntactic analysis is used to identify not only noun phrases, but also verb phrases, adjectival phrases and more, according to user’s needs.

Concepts are minimum fragments of text which refer to **objects** or **ideas**. The service can be used to extract objects and ideas from any type of text, either high-quality ones (news, legislation...) or colloquial ones (forums, blogs, chats, social media...).

NaturalExtractor uses Deep Linguistic Analysis based on grammars, which allows for a rich variety of concept detection adapted according to user’s needs:

- Simple concepts: “checkingaccount”, “Corporate policy”
- Compound or nested concepts: “Company’s checking account”
- Combinations of the above: “account”, “checking account”, “checking account of the bank”, “bank”.

The linguistic analysis also makes it possible to extract concepts based on syntactic phrases:

- Noun phrases: “Are criteria used for **individual director** performance evaluation disclosed?”
- Adjectival phrases: “Are **criteria** used for individual director performance evaluation **disclosed**?”
- Verb phrases: “**Are** criteria used for individual director performance evaluation **disclosed**?”

The results from NaturalExtractor for few examples of such questions are,

Table 5.4 Result from Natural Extractor

QUESTIONNAIRE	KEYWORDS
Are the type of material transactions that must be approved by the board disclosed?	Materials transaction approval
Was appraisal of board performance conducted?	Appraisal of board performance details

Statement of Company's philosophy on code of governance.	Corporate governance philosophy
Do the criteria used for evaluating board performance include financial measures such as the company's share price performance or total shareholder return?	Criteria for board evaluation, financial measures
Are criteria used for individual director performance evaluation disclosed?	Criteria for individual director evaluation

5.5 SCORE CARD PREPARATION

The score card contains 115 features in which each one has an assigned score. The score of the features falls under three categories, each has different scores such as 1,2 and 5. It takes the annual report and dictionary as input in order to perform matching.

N-gram matching technique is used for the comparison of features with annual reports. When the match is found, the corresponding score is assigned to that particular feature. Otherwise the feature is assigned with null value.

N-gram models are widely used in statistical natural language processing. For parsing, words are modelled such that each n-gram is composed of n words. For sequences of words, the trigrams that can be generated from "Statement of Company's philosophy on code of governance." are "# the company", "statement of company", "company philosophy #", "code of governance", "philosophy on #" and "statement of #". Initial step is to pre-process the strings to remove spaces, most simply collapse whitespace to a single space while preserving paragraph marks. Punctuation is also commonly reduced or removed by pre-processing

5.6 FEATURE SELECTION

5.6.1 T-Test Evaluation

A t-test is used to compare whether the two groups have different average values. It can be done by using the MATLAB built-in function called `mattest()`. The general form of the function is given below.

```
PValues = mattest(DataX, DataY)
[PValues, TScores] = mattest(DataX, DataY)
```

DataMatrix object is a matrix of feature values where each row corresponds to a feature and each column corresponds to a company. DataX and DataY must have the same number of rows and are assumed to be normally distributed in each class with equal variances.

Where DataX contains data from well growing companies and DataY contains data from a low performing companies. For example, DataX could be feature values from TCS, Infosys, and DataY could be feature values from Zenith Birla and KingFisher Airlines.

The output from the above function is a matrix which contains a column vector of p-values and a column vector of t-scores. Then the p-values column vector is arranged in descending order in order to find the average similarity between the features. Now the top p-values which satisfies the threshold are taken as a new feature set. So that the common features are extracted from the whole 115 features.

5.7 ASSOCIATIVE CLASSIFICATION

Association classification is a recent and rewarding technique that combines the methodology of association and classification. This generally involves two stages:

1. Generate class association rules from a training data set.
2. Classify the test data set into predefined class labels.

In short, Association classification utilizes the association rule discovery techniques to construct classification systems.

An Association Rule is a rule of the form

$$\textit{cot and bed} \rightarrow \textit{pillow}$$

where 'cot and bed' is called the rule body and pillow the head of the rule. It associates the rule body with its head. The following example expresses the fact that people who buy cot and bed are likely to buy pillow too.

An association rule can be defined as: Let D be a database consisting of one table over n attributes $\{a_1, a_2, \dots, a_n\}$. Let this table contain k instances. Let d be a database record. d satisfies an item set $X \subseteq \{a_1, a_2, \dots, a_n\}$ if $X \subseteq d$. An association rule is an implication $X \rightarrow Y$ where $X, Y \subseteq \{a_1, a_2, \dots, a_n\}$. The use of association rules for classification is restricted to problems where the instances can only belong to a discrete number of classes. The reason is that association rule mining is only possible for nominal attributes. However, association rules in their general form cannot be used directly. Their definition has to be restricted. The head Y of an arbitrary association rule $X \rightarrow Y$ is a disjunction of items. Every item which is not present in the rule body may occur in the head of the rule. The rules that are capable of assigning a class membership can be used as rules for classification. Therefore the head Y of a class association rule $X \rightarrow Y$ is restricted to one item. The attribute of this attribute-value-pair has to be the class attribute.

According to this, a class association rule is of the form $X \rightarrow a_i$ where a_i is the class attribute and $X \subseteq \{a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n\}$

5.7.1 Steps in Associative Classification Rule Mining

The steps in associative classification can be summarized as follows:

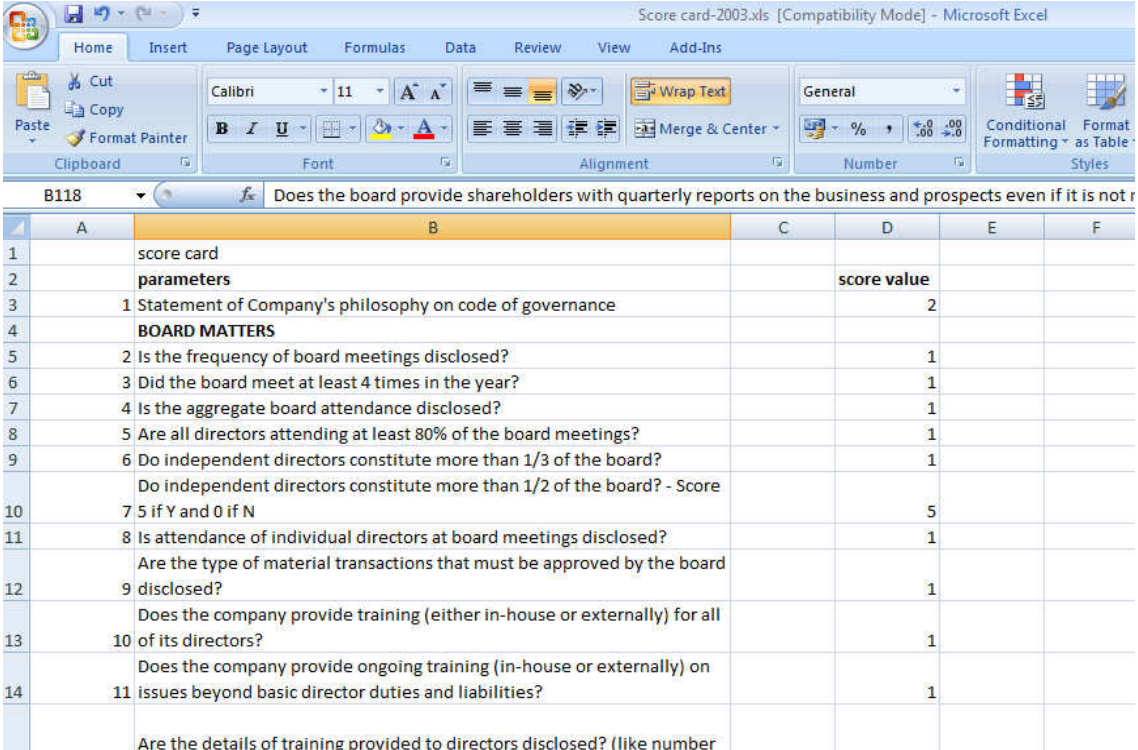
- The discovery of all frequent rule items from a training dataset which yields class association rules.
- Ranking of the Class Association Rules (CAR) and pruning them based on the measures of interestingness (Support, Confidence) and their threshold values.
- The selection of one subset of CARs to form the classifier.
- Measuring the quality of the derived classifier on test data objects.

CHAPTER 6

EXPERIMENTAL RESULTS

6.1 SCORE CARD FORMATION

The sample score card with 115 features and their assigned score is given in Figure 6.1.



The screenshot shows a Microsoft Excel spreadsheet titled "Score card-2003.xls [Compatibility Mode] - Microsoft Excel". The spreadsheet contains a table with 6 columns (A to F) and 15 rows. The table is a score card with 115 features and their assigned scores. The features are listed in column A, and the scores are listed in column D. The table is titled "score card" and "parameters" in row 1. The features are listed in rows 2 to 15, and the scores are listed in rows 2 to 15. The table is titled "score card" and "parameters" in row 1. The features are listed in rows 2 to 15, and the scores are listed in rows 2 to 15.

	A	B	C	D	E	F
1	score card					
2	parameters			score value		
3	1 Statement of Company's philosophy on code of governance			2		
4	BOARD MATTERS					
5	2 Is the frequency of board meetings disclosed?			1		
6	3 Did the board meet at least 4 times in the year?			1		
7	4 Is the aggregate board attendance disclosed?			1		
8	5 Are all directors attending at least 80% of the board meetings?			1		
9	6 Do independent directors constitute more than 1/3 of the board?			1		
10	7 5 if Y and 0 if N			5		
11	8 Is attendance of individual directors at board meetings disclosed?			1		
12	9 Are the type of material transactions that must be approved by the board disclosed?			1		
13	10 Does the company provide training (either in-house or externally) for all of its directors?			1		
14	11 Does the company provide ongoing training (in-house or externally) on issues beyond basic director duties and liabilities?			1		
	Are the details of training provided to directors disclosed? (like number					

Figure 6.1 Score Card Representation

The score card generated for companies such as CTS, WIPRO, TCS, HCL, KingFisher Airlines, ITC and Zenith Brila is given in Figure 6.1.1

Book1.xlsx - Microsoft Excel

Home

Insert

Page Layout

Formulas

Data

Review

View

Add-Ins

Courier New

10.5

A

A

B

I

U

Font

Alignment

General

%

Number

Conditional Formatting

Format as Table

Cell Styles

Styles

Insert

Delete

Format

Cells

Sort & Filter

Find & Select

Editing

S14

1

	E	F	G	H	I	J	K	L	M	N	O	P	Q	R		
1	2007-08				2008-09				2009-10				2010-11			
2	WIPRO	TCS	HCL	CTS	WIPRO	TCS	HCL	CTS	WIPRO	TCS	HCL	CTS	WIPRO	TCS	HCL	
3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	
4																
5	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	
6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
7	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	
8	0	1	1	0	0	1	1	0	1	1	1	0	0	1	1	
9	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
11	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	1	1	1	0	1	0	1	0	1	0	1	0	0	1	1	

Sheet1

Sheet2

Sheet3

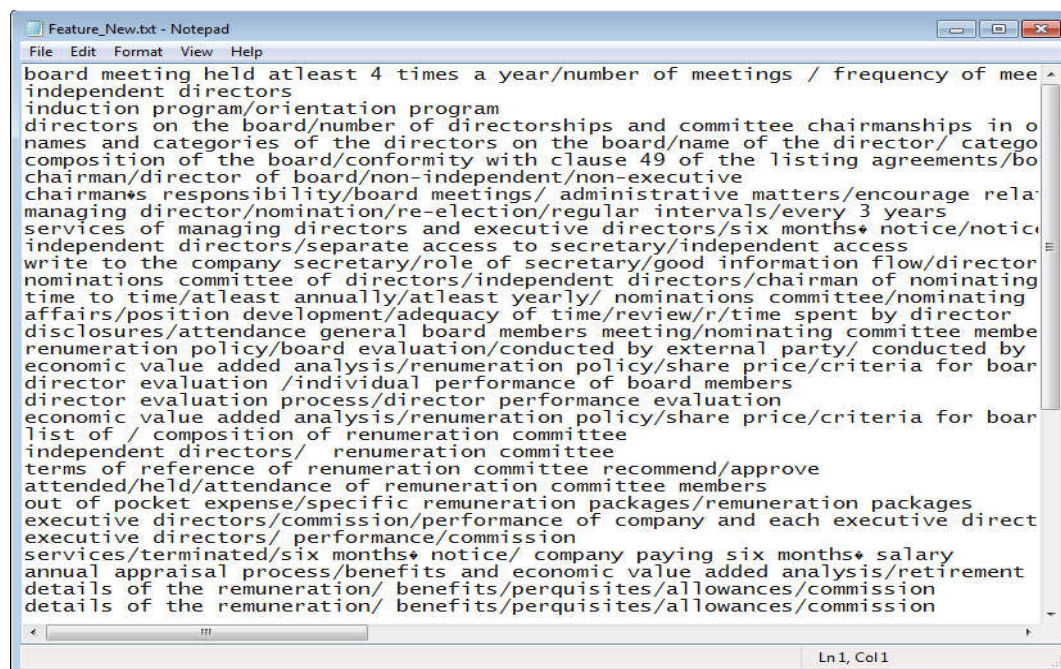
Ready

100%

Figure 6.1.1 Score Card

6.2 Feature Selection

Among 115 features, 76 features are selected based on T-test evaluation and are listed in Figure 6.2



```

board meeting held atleast 4 times a year/number of meetings / frequency of mee
independent directors
induction program/orientation program
directors on the board/number of directorships and committee chairmanships in o
names and categories of the directors on the board/name of the director/ catego
composition of the board/conformity with clause 49 of the listing agreements/bo
chairman/director of board/non-independent/non-executive
chairman*s responsibility/board meetings/ administrative matters/encourage rela
managing director/nomination/re-election/regular intervals/every 3 years
services of managing directors and executive directors/six months* notice/notic
independent directors/separate access to secretary/independent access
write to the company secretary/role of secretary/good information flow/director
nominations committee of directors/independent directors/chairman of nominating
time to time/atleast annually/atleast yearly/ nominations committee/nominating
affairs/position development/adequacy of time/review/r/time spent by director
disclosures/attendance general board members meeting/nominating committee membe
remuneration policy/board evaluation/conducted by external party/ conducted by
economic value added analysis/remuneration policy/share price/criteria for boar
director evaluation /individual performance of board members
director evaluation process/director performance evaluation
economic value added analysis/remuneration policy/share price/criteria for boar
list of / composition of remuneration committee
independent directors/ remuneration committee
terms of reference of remuneration committee recommend/approve
attended/held/attendance of remuneration committee members
out of pocket expense/specific remuneration packages/remuneration packages
executive directors/commission/performance of company and each executive direct
executive directors/ performance/commission
services/terminated/six months* notice/ company paying six months* salary
annual appraisal process/benefits and economic value added analysis/retirement
details of the remuneration/ benefits/perquisites/allowances/commission
details of the remuneration/ benefits/perquisites/allowances/commission

```

Figure 6.2 Feature Selection

6.3 Associative Classification

By using the above selected features, training data is given and then the association rules are generated. Now the testing is done from those rules.

[illegible]

Figure 6.3 Classification Result

6.4 Summary Of Dataset

The Annual reports from the firms such as TCS, Wipro, Zenith Birla and KingFisherAirlines,etc., are taken. The details on the count of documents are given in Table 6.1

Table 6.1 Summary Of Dataset

Name Of The Company	Number of Documents
CTS	7
TCS	7

Name Of The Company	Number of Documents
Wipro	5
HCL	7
Zenith Birla	6
KingFisherAirLines	5
M-Phasis	6
ITC	7
Total Number Of Documents	50

The first column represents the name of the company. The second column gives the number of documents in each company. The last row represents the total number of documents.

6.5 ACCURACY COMPUTATION

Accuracy measures the ability of the classifier to correctly classify unlabeled data. It is the ratio of the number of correctly classified data over the total number of given transactions in the test dataset.

$$\text{Accuracy} = \frac{\text{Number of correctly predicted test data}}{\text{total number of test data}}$$

Table 6.2 Accuracy computation

	Accuracy(in percentage)	Number Of Rules
Without Feature Selection	75.0	16
With Feature Selection	87.5	16

The second row gives the accuracy calculated for 115 features. The third row gives the accuracy calculated for 76 features(which is derived from 115 features using T-test).

From the above tabulated data, accuracy computed before feature selection is lesser than the accuracy with feature selection. Thus feature selection by T-test gives enhanced results.

CHAPTER 7

CONCLUSION

Every day, news of financial statement fraud is adversely affecting the economy worldwide. Prediction of financial statement fraud would be of great value to the organizations throughout the world. Considering the need of such a mechanism, a data mining framework is to be employed for prediction of financial statement fraud.

Thus our project aims to expand the understanding of how deceivers use language differently than truth tellers, particularly in high-stakes, real-world environments such as financial markets. It takes a radically different approach to data processing and analytics, by combining with sophisticated analytics to predict fraudulent companies. This facilitates customers to identify the status of companies from the annual reports published by them.

BIBLIOGRAPHY

- [1]S. Kotsiantis,Euaggelos koumanakos,Dimitris Tzelepis and Vasilis Tampakas.,“Predicting fraudulent financial statements with machine learning techniques”, International Journal of Computational Intelligence, Vol. 3, No. 2, pp.104-110, 2006.
- [2]Johan L. Perols.,“Detecting financial statement fraud”,International Journal of Computer Applications,2008.
- [3]Rajan Gupta, Nasib Singh Gill., “Prevention and Detection of Financial Statement fraud”, International Journal of Advanced Computer Science and Applications, Vol. 3, No. 8, 2012.
- [4]Efsthios Kirkos, Charalambos Spathis, Yannis Manolopoulos.,“Detection Of Fraudulent Financial Statements Through The Use Of Data Mining Techniques”,2nd International Conference on Enterprise Systems and Accounting,2005.
- [5]Sotiris Kotsiantis,“Multi-Instance Learning For Predicting Fraudulent Financial Statements”,Third International Conference In Convergence and Hybrid Information Technology,, Vol. 1,2008.
- [6]Anuj Sharma,Prabin Kumar Panigrahi.,“A Review Of Financial Accounting Fraud Detection Based On Data Mining Technique”, International Journal of Computer Applications (0975 – 8887)Volume 39– No.1, 2013.
- [7]O. Persons, "Using financial statement data to identify factors associated with fraudulent financing reporting," Journal of Applied Business Research,vol. 11, pp. 38-46, 1995.
- [8]C. Spathis, "Detecting false financial statements usingpublished, data: some evidence from Greece,"Managerial Auditing Journal, vol. 17, no. 4, pp.179-191,2002.
- [9]J. W. Lin, M. I. Hwang, and J. D. Becker, "A fuzzyneural network for assessing the risk of fraudulentfinancial reporting," Managerial Auditing Journal,Vol. 18, pp. 657-665,2003.