

On The Use Of Text Classification Methods For Text Summarisation

Thesis submitted in accordance with the requirements of
the University of Liverpool for the degree of Doctor in Philosophy
by
Matias Fernando Garcia-Constantino

July 2013

Dedication

To my parents and my girlfriend

Acknowledgement

First and foremost, I am very grateful to my first supervisor Prof. Frans Coenen, for all his patience, constant support, excellent advice and encouragement throughout my four-year study. Without him, the successful completion of this thesis would not have been possible. It has been a privilege to have worked with him.

I also want to thank my second supervisors from the Veterinary Science School, Dr. Alan Radford and Dr. P-J Noble, for their support and advice regarding the Veterinary Science data used in my research. Special thanks also to Dr. Christian Setzkorn for his useful opinions and advice.

I would also like to acknowledge my family and friends for all their love, motivation and support that has helped me to carry on through my PhD studies. Particular thanks go to: my parents Patricia Constantino Casas and Matías García Ruiz have always supported and encouraged me in all my plans and decisions; my uncle Dr. Fernando Constantino Casas and my friends Dr. Simon Holgate, Dr. Luis Carlos Molina Félix, Dr. Stalin Muñoz Gutiérrez and Dr. Omar Baqueiro Espinosa for all their advice and support. I would also like to thank: Dr. Rolando Medellín Gasque, which was also doing his PhD but one year ahead, for his friendship, patience and support; my colleague Dr. Stephanie Chua for all her support; my colleague Agneau Belanyek for his friendship and support; and my girlfriend Sarah Cullen, for all her love, support and inspiration.

Last but not least, I want to extend my gratitude to the Consejo Nacional de Ciencia y Tecnología (CONACYT) for the financial assistance they provided for my PhD studies.

Abstract

This thesis describes research work undertaken in the fields of text and questionnaire mining. More specifically, the research work is directed at the use of text classification techniques for the purpose of summarising the free text part of questionnaires. In this thesis text summarisation is conceived of as a form of text classification in that the classes assigned to text documents can be viewed as an indication (summarisation) of the main ideas of the original free text but in a coherent and reduced form. The reason for considering this type of summary is because summarising unstructured free text, such as that found in questionnaires, is not deemed to be effective using conventional text summarisation techniques. Four approaches are described in the context of the classification summarisation of free text from different sources, focused on the free text part of questionnaires. The first approach considers the use of standard classification techniques for text summarisation and was motivated by the desire to establish a benchmark with which the more specialised summarisation classification techniques presented later in this thesis could be compared. The second approach, called Classifier Generation Using Secondary Data (CGUSD), addresses the case when the available data is not considered sufficient for training purposes (or possibly because no data is available at all). The third approach, called Semi-Automated Rule Summarisation Extraction Tool (SARSET), presents a semi-automated classification technique to support document summarisation classification in which there is more involvement by the domain experts in the classifier generation process, the idea was that this might serve to produce more effective summaries. The fourth is a hierarchical summarisation classification approach which assumes that text summarisation can be achieved using a classification approach whereby several class labels can be associated with documents which then constitute the summarisation. For evaluation purposes three types of text were considered: (i) questionnaire free text, (ii) text from medical abstracts and (iii) text from news stories.

Contents

Dedication	i
Acknowledgement	ii
Abstract	iii
Contents	iv
List of Figures	x
List of Tables	xi
1 Introduction	1
1.1 Overview	1
1.2 Motivation	3
1.3 Research question and issues	5
1.4 Problem definition	6
1.5 Research Methodology	6
1.6 Contributions	7
1.7 Organisation of the thesis	8
1.8 Publications	9
1.9 Summary	11
2 Literature Review	12
2.1 Introduction	12
2.2 Data Mining and the Knowledge Discovery in Databases (KDD) process	12
2.3 Questionnaire Data Mining (QDM)	14
2.3.1 Overview of QDM approaches	15
2.3.2 QDM approaches directed at closed-ended questions (tabular data)	16
2.3.3 QDM approaches directed at open-ended questions (free text) . .	17
2.3.4 QDM approaches directed at both closed-ended and open-ended questions (tabular data and free text)	19
2.4 Text Classification	20

2.4.1	Feature Selection in Text Classification	21
2.4.1.1	Information Gain (IG)	21
2.4.1.2	Chi-squared (χ^2) statistic	22
2.4.1.3	Correlation Feature Selection (CFS)	23
2.4.1.4	Feature Weighting with Term Frequency-Inverse Document Frequency (TF-IDF)	23
2.4.1.5	Feature selection methods used	24
2.4.2	Relevant Text Classification algorithms	25
2.4.2.1	Bayesian Classifiers	25
2.4.2.2	Decision Trees	26
2.4.2.3	Rule Learners	26
2.4.2.4	Nearest Neighbour Techniques	27
2.4.2.5	Support Vector Machines (SVM)	27
2.4.3	Evaluation measures for Text Classification	28
2.4.3.1	Confusion Matrix	28
2.4.3.2	Accuracy	28
2.4.3.3	Precision	29
2.4.3.4	Recall/Sensitivity	29
2.4.3.5	Specificity	29
2.4.3.6	Area Under the ROC Curve (AUC)	29
2.5	Text Summarisation	30
2.5.1	Categorisation of Text Summarisation techniques	30
2.5.2	Relevant Text Summarisation approaches	33
2.5.2.1	Techniques based on lexical chains	33
2.5.2.2	Techniques based on sentence extraction and ranking	34
2.5.2.3	Other techniques	35
2.5.3	Text Summarisation approaches that use Text Classification methods	36
2.5.4	Evaluation measures for Text Summarisation	36
2.6	Summary	37
3	Evaluation Data Sets and Data Preprocessing	38
3.1	Introduction	38
3.2	Data Preprocessing	40
3.2.1	Tabular Data	40
3.2.2	Free Text	41
3.3	Small Animal Veterinary Surveillance Network (SAVSNET)	43
3.3.1	Description of the SAVSNET data set	43
3.3.1.1	SAVSNET-840-4-FT	44
3.3.1.2	SAVSNET-840-4-TD+FT	44

3.3.1.3	SAVSNET-971-3-FT	45
3.3.1.4	SAVSNET-971-3-TD+FT	45
3.3.1.5	SAVSNET-917-4H	45
3.3.2	Preprocessing of the SAVSNET data set	47
3.4	OHSUMED	53
3.4.1	Description of the OHSUMED data set	54
3.4.1.1	OHSUMED-CA-3187-3H (Cardiovascular Abnormalities)	55
3.4.1.2	OHSUMED-AD-3393-3H (Animal Diseases)	55
3.4.2	Preprocessing of the OHSUMED data set	56
3.5	Reuters-21578	59
3.5.1	Description of the Reuters-21578 data set	59
3.5.1.1	Reuters-21578-LOC-2327-2H	60
3.5.1.2	Reuters-21578-COM-2327-2H	61
3.5.2	Preprocessing of the Reuters-21578 data set	61
3.6	Summary	63
4	Using Standard Classification Techniques for Text Summarisation	67
4.1	Introduction	67
4.2	Methodology	68
4.2.1	Classification	68
4.2.2	Summary Generation	69
4.3	Experiments and Results	70
4.3.1	SAVSNET-840-4	71
4.3.2	SAVSNET-971-3	75
4.3.3	SAVSNET-917-4H	77
4.3.3.1	SAVSNET-917-1L	77
4.3.3.2	SAVSNET-917-2L	77
4.3.3.3	SAVSNET-917-3L	79
4.3.3.4	SAVSNET-917-4L	79
4.3.4	OHSUMED-CA-3187-3H	81
4.3.4.1	OHSUMED-CA-3187-1L	81
4.3.4.2	OHSUMED-CA-2570-2L	81
4.3.4.3	OHSUMED-CA-834-3L	83
4.3.5	OHSUMED-AD-3393-3H	85
4.3.5.1	OHSUMED-AD-3393-1L	85
4.3.5.2	OHSUMED-AD-569-2L	85
4.3.5.3	OHSUMED-AD-292-3L	86
4.3.6	Reuters-21578-LOC-2327-2H	88
4.3.6.1	Reuters-21578-LOC-2327-1L	88
4.3.6.2	Reuters-21578-LOC-2327-2L	88

4.3.7	Reuters-21578-COM-2327-2H	90
4.3.7.1	Reuters-21578-COM-2327-1L	90
4.3.7.2	Reuters-21578-COM-2327-2L	90
4.3.8	Evaluation Summary	92
4.4	Text Summarisation	94
4.5	Discussion	105
4.6	Summary	106
5	Classifier Generation Using Secondary Data (CGUSD) for Text Sum-	
	marisation	107
5.1	Introduction	107
5.2	The Classifier Generation Using Secondary Data (CGUSD) Methodology	108
5.2.1	Secondary data set generation	110
5.2.2	Classification	112
5.3	Secondary data sets used	113
5.3.1	MEDLINE (Medical Literature Analysis and Retrieval System	
	Online)	113
5.3.1.1	SAVSNET	114
5.3.1.2	OHSUMED	114
5.3.2	RCV1 (Reuters Corpus Volume 1)	115
5.3.2.1	Reuters-21578	115
5.4	Experiments and Results	115
5.4.1	SAVSNET-840-4-FT	117
5.4.2	SAVSNET-971-3-FT	118
5.4.3	SAVSNET-917	119
5.4.3.1	SAVSNET-917-1L	119
5.4.4	OHSUMED-CA-3187	119
5.4.4.1	OHSUMED-CA-3187-1L	120
5.4.4.2	OHSUMED-CA-2570-2L	120
5.4.4.3	OHSUMED-CA-834-3L	122
5.4.5	OHSUMED-AD-3393	123
5.4.5.1	OHSUMED-AD-3393-1L	123
5.4.5.2	OHSUMED-AD-569-2L	123
5.4.5.3	OHSUMED-AD-292-3L	124
5.4.6	Reuters-21578-LOC-2327-2H	125
5.4.6.1	Reuters-21578-LOC-2327-1L	125
5.4.6.2	Reuters-21578-LOC-2327-2L	125
5.4.7	Reuters-21578-COM-2327-2H	127
5.4.7.1	Reuters-21578-COM-2327-1L	127
5.5	Discussion	128

5.6	Summary	129
6	Using a Semi-Automated Rule Summarisation Extraction Tool (SARSET) for Text Summarisation	131
6.1	Introduction	131
6.2	The SARSET Methodology	133
6.2.1	Problem definition	133
6.2.2	Classifier Generation Using SARSET (Semi-Automated Rule Summarisation Extraction Tool)	134
6.2.3	Phrase identification and generation of phrase variations (Step 1)	134
6.2.4	Identification of questionnaires covered by identified phrases (Step 2)	136
6.2.5	Rule generation (Steps 3 and 4)	137
6.2.6	Continuation of the process or exit (Step 5)	139
6.2.7	Applying classification rules to unseen documents	139
6.3	Experiments and Results	140
6.3.1	SAVSNET-840-4-FT	141
6.3.2	SAVSNET-971-3-FT	142
6.3.3	SAVSNET-917	142
6.3.3.1	SAVSNET-917-1L	142
6.3.3.2	SAVSNET-917-2L	143
6.3.3.3	SAVSNET-917-3L	143
6.3.3.4	SAVSNET-917-4L	144
6.3.4	OHSUMED-CA-3187	144
6.3.4.1	OHSUMED-CA-3187-1L	144
6.3.4.2	OHSUMED-CA-2570-2L	145
6.3.4.3	OHSUMED-CA-834-3L	145
6.3.5	OHSUMED-AD-3393	145
6.3.5.1	OHSUMED-AD-3393-1L	146
6.3.5.2	OHSUMED-AD-569-2L	146
6.3.5.3	OHSUMED-AD-292-3L	146
6.3.6	Reuters-21578-LOC-2327-2H	147
6.3.6.1	Reuters-21578-LOC-2327-1L	147
6.3.6.2	Reuters-21578-LOC-2327-2L	147
6.3.7	Reuters-21578-COM-2327-2H	148
6.3.7.1	Reuters-21578-COM-2327-1L	148
6.3.7.2	Reuters-21578-COM-2327-2L	148
6.4	Discussion	149
6.5	Summary	152

7	Text Summarisation Using Hierarchical Text Classification	154
7.1	Introduction	154
7.2	Methodology	157
7.2.1	Problem definition	157
7.2.2	Hierarchical Classification	158
7.2.3	Summary Generation	160
7.3	Hierarchical data sets used	160
7.3.1	SAVSNET-917-4H	160
7.3.2	OHSUMED-CA-3187-3H	161
7.3.3	OHSUMED-AD-3393-3H	162
7.3.4	Reuters-21578	162
7.3.4.1	Reuters-21578-LOC-2327-2H	163
7.3.4.2	Reuters-21578-COM-2327-2H	163
7.4	Experiments and Results	164
7.4.1	SAVSNET-917-4H	169
7.4.2	OHSUMED-CA-3187-3H	169
7.4.3	OHSUMED-AD-3393-3H	170
7.4.4	Reuters-21578-LOC-2327-2H	170
7.4.5	Reuters-21578-COM-2327-2H	170
7.4.6	Text Summarisation	171
7.4.6.1	SAVSNET	171
7.4.6.2	OHSUMED	171
7.4.6.3	Reuters-21578	172
7.5	Discussion	172
7.6	Summary	181
8	Conclusions and Future Work	182
8.1	Summary	182
8.2	Main Findings and Contributions	185
8.3	Future Directions	188
	Bibliography	191
A	Distribution of records per class with respect to the evaluation data sets used	201
B	Additional Experimental Results	211
C	Published Work	222

List of Figures

2.1	Goals of Data Mining	13
2.2	Questionnaire Data Mining relevant approaches.	16
2.3	Text Classification.	20
2.4	Input factors to be considered for text summarisation	31
2.5	Purpose factors to be considered for text summarisation	31
2.6	Output factors to be considered for text summarisation	32
3.1	Example fragment of raw questionnaire data in CSV format.	48
3.2	Example fragment of preprocessed free text merged with tabular attributes.	52
3.3	Example fragment of tabular questionnaire data in CSV format.	54
3.4	Example of raw OHSUMED data.	57
3.5	Example of partially preprocessed OHSUMED data.	59
3.6	Example fragment of raw Reuters-21578 data.	64
3.7	Two example preprocessed Reuters-21578 records.	65
4.1	Standard Classification Technique for Text Summarisation.	73
5.1	Classifier Generation Using Secondary Data (CGUSD) for Text Summarisation.	109
6.1	The SARSET methodology.	135
6.2	Main window of SARSET.	136
6.3	Validation window of SARSET.	137
6.4	Inspection of documents.	138
7.1	Hierarchy of classes.	158
7.2	Single-parent hierarchy	158
7.3	Multi-parent hierarchy	158

List of Tables

2.1	Example of a binary confusion matrix	28
3.1	Number of records per class in SAVSNET-840-4-FT.	44
3.2	Number of records per class in SAVSNET-971-3-FT.	45
3.3	Number of records per class in SAVSNET-917-1L.	46
3.4	Number of records per class in SAVSNET-917-2L.	46
3.5	Number of records per class in SAVSNET-917-3L.	46
3.6	Number of records per class in SAVSNET-917-4L.	46
3.7	Example fragment of raw questionnaire data in tabular form.	49
3.8	Number of classes per level in the OHSUMED-CA-3187-3H hierarchy. .	55
3.9	Number of classes per level in the OHSUMED-AD-3393-3H hierarchy. .	56
3.10	Field definitions for OHSUMED data set.	58
3.11	Number of classes per level in the Reuters-21578-LOC-2327-2H hierarchy.	61
3.12	Number of classes per level in the Reuters-21578-COM-2327-2H hierarchy.	61
3.13	Comparison of data sets and their usage in relation to the proposed methods.	66
4.1	Example of rules for generating summaries.	70
4.2	Classification results for the SAVSNET-840-4 data set with Chi-squared.	74
4.3	Classification results for the SAVSNET-840-4 data set with CFS.	74
4.4	Classification results for the SAVSNET-971-3 data set with Chi-squared.	76
4.5	Classification results for the SAVSNET-971-3 data set with CFS.	76
4.6	Classification results for the SAVSNET-917-1L data set with Chi-squared and CFS.	78
4.7	Classification results for the SAVSNET-917-2L data set with Chi-squared and CFS.	78
4.8	Classification results for the SAVSNET-917-3L data set with Chi-squared and CFS.	80
4.9	Classification results for the SAVSNET-917-4L data set with Chi-squared and CFS.	80
4.10	Classification results for the OHSUMED-CA-3187-1L data set with Chi- squared and CFS.	82

4.11	Classification results for the OHSUMED-CA-2570-2L data set with Chi-squared and CFS.	82
4.12	Classification results for the OHSUMED-CA-834-3L data set with Chi-squared and CFS.	84
4.13	Classification results for the OHSUMED-AD-3393-1L data set with Chi-squared and CFS.	84
4.14	Classification results for the OHSUMED-AD-569-2L data set with Chi-squared and CFS.	87
4.15	Classification results for the OHSUMED-AD-292-3L data set with Chi-squared and CFS.	87
4.16	Classification results for the Reuters-21578-LOC-2327-1L data set with Chi-squared and CFS.	89
4.17	Classification results for the Reuters-21578-LOC-2327-2L data set with Chi-squared and CFS.	89
4.18	Classification results for the Reuters-21578-COM-2327-1L data set with Chi-squared and CFS.	91
4.19	Classification results for the Reuters-21578-COM-2327-2L data set with Chi-squared and CFS.	91
4.20	Best classification techniques and results.	93
4.21	Examples of summaries generated using standard classification techniques.	96
5.1	Statistical details for primary and secondary data sets used.	117
5.2	SAVSNET-840-4-FT primary and secondary data sets ($k = 419$).	117
5.3	Classification results for the SAVSNET-840-4-FT data set.	118
5.4	SAVSNET-971-3-FT primary and secondary data sets ($k = 429$).	118
5.5	Classification results for the SAVSNET-971-3-FT data set.	119
5.6	SAVSNET-917-1L primary and secondary data sets ($k = 429$).	119
5.7	Classification results for the SAVSNET-917-1L data set.	120
5.8	OHSUMED-CA-3187-1L primary and secondary data sets ($k = 2,339$).	120
5.9	Classification results for the OHSUMED-CA-3187-1L data set.	120
5.10	OHSUMED-CA-2570-2L primary and secondary data sets ($k = 386$).	121
5.11	Classification results for the OHSUMED-CA-2570-2L data set.	122
5.12	OHSUMED-CA-834-3L primary and secondary data sets ($k = 446$).	122
5.13	Classification results for the OHSUMED-CA-834-3L data set.	122
5.14	Classification results for the OHSUMED-AD-3393-1L data set.	123
5.15	Classification results for the OHSUMED-AD-569-2L data set.	124
5.16	OHSUMED-AD-292-3L primary and secondary data sets ($k = 292$).	124
5.17	Classification results for the OHSUMED-AD-292-3L data set.	125
5.18	Reuters-21578-LOC-2327-1L primary and secondary data sets ($k = 214$).	126
5.19	Classification results for the Reuters-21578-LOC-2327-1L data set.	126

5.20	Classification results for the Reuters-21578-LOC-2327-2L data set. . . .	127
5.21	Reuters-21578-COM-2327-1L primary and secondary data sets ($k = 600$). . . .	127
5.22	Classification results for the Reuters-21578-COM-2327-1L data set. . . .	128
5.23	Best classification techniques and results.	129
6.1	Classification results for the SAVSNET-840-4-FT data set.	142
6.2	Classification results for the SAVSNET-971-3-FT data set.	142
6.3	Classification results for the SAVSNET-917-1L data set.	143
6.4	Classification results for the SAVSNET-917-2L data set.	143
6.5	Classification results for the SAVSNET-917-3L data set.	143
6.6	Classification results for the SAVSNET-917-4L data set.	144
6.7	Classification results for the OHSUMED-CA-3187-1L data set.	144
6.8	Classification results for the OHSUMED-CA-2570-2L data set.	145
6.9	Classification results for the OHSUMED-CA-834-3L data set.	145
6.10	Classification results for the OHSUMED-AD-3393-1L data set.	146
6.11	Classification results for the OHSUMED-AD-569-2L data set.	146
6.12	Classification results for the OHSUMED-AD-292-3L data set.	147
6.13	Classification results for the Reuters-21578-LOC-2327-1L data set. . . .	147
6.14	Classification results for the Reuters-21578-LOC-2327-2L data set. . . .	148
6.15	Classification results for the Reuters-21578-COM-2327-1L data set. . . .	148
6.16	Classification results for the Reuters-21578-COM-2327-2L data set. . . .	149
6.17	Best classification results.	150
7.1	Classification results for SAVSNET-917-4H using SMO.	166
7.2	Classification results for OHSUMED-CA-3187-3H using SMO.	166
7.3	Averaged accuracy and AUC results for SAVSNET-917-4H using SMO, C4.5 and RIPPER.	168
7.4	Averaged accuracy and AUC results for OHSUMED-CA-3187-3H using SMO, C4.5 and RIPPER.	168
7.5	Averaged accuracy and AUC results for Reuters-21578-LOC-2327-2H us- ing SMO, C4.5 and RIPPER.	168
7.6	Averaged accuracy and AUC results for Reuters-21578-COM-2327-2H using SMO, C4.5 and RIPPER.	168
7.7	Best classification techniques and results.	173
7.8	Overall classification results for the proposed approaches.	177
A.1	Number of records per class in SAVSNET-840-4 and in SAVSNET-840- 4-TD+FT ($k = 352$).	201
A.2	Number of records per class in SAVSNET-971-3 and in SAVSNET-971- 3-TD+FT ($k = 586$).	202

A.3	Number of records per class in SAVSNET-917-1L ($k = 536$).	202
A.4	Number of records per class in SAVSNET-917-2L ($k = 604$).	202
A.5	Number of records per class in SAVSNET-917-3L ($k = 573$).	202
A.6	Number of records per class in SAVSNET-917-4L ($k = 411$).	202
A.7	Number of records per class in OHSUMED-CA-3187-1L ($k = 2,339$).	203
A.8	Number of records per class in OHSUMED-CA-2570-2L ($k = 525$).	203
A.9	Number of records per class in OHSUMED-CA-834-3L ($k = 299$).	203
A.10	Number of records per class in OHSUMED-AD-3393-1L ($k = 2,112$).	204
A.11	Number of records per class in OHSUMED-AD-569-2L ($k = 194$).	205
A.12	Number of records per class in OHSUMED-AD-292-3L ($k = 133$).	205
A.13	Number of records per class in Reuters-21578-LOC-2327-1L ($k = 1,021$).	206
A.14	Number of records per class in Reuters-21578-COM-2327-1L ($k = 966$).	206
A.15	Number of records per class in Reuters-21578-LOC-2327-2L ($k = 743$).	207
A.16	Number of records per class in Reuters-21578-COM-2327-2L ($k = 508$).	209
B.1	Classification results for SAVSNET-917-4H using SMO.	212
B.2	Classification results for SAVSNET-917-4H using C4.5.	212
B.3	Classification results for SAVSNET-917-4H using RIPPER.	213
B.4	Classification results for OHSUMED-CA-3187-3H using SMO.	213
B.5	Classification results for OHSUMED-CA-3187-3H using C4.5.	214
B.6	Classification results for OHSUMED-CA-3187-3H using RIPPER.	214
B.7	Classification results for OHSUMED-AD-3393-3H using SMO.	215
B.8	Classification results for OHSUMED-AD-3393-3H using C4.5.	215
B.9	Classification results for OHSUMED-AD-3393-3H using RIPPER.	216
B.10	Classification results for Reuters-21578-LOC-2327-2H using SMO.	217
B.11	Classification results for Reuters-21578-LOC-2327-2H using C4.5.	218
B.12	Classification results for Reuters-21578-LOC-2327-2H using RIPPER.	219
B.13	Classification results for Reuters-21578-COM-2327-2H using SMO.	220
B.14	Classification results for Reuters-21578-COM-2327-2H using C4.5.	220
B.15	Classification results for Reuters-21578-COM-2327-2H using RIPPER.	221

Chapter 1

Introduction

1.1 Overview

In the last decades there has been a dramatic increase in the number of computer users, mainly because of the popularization of the Internet and the feasibility of automating office processes using dedicated software. This has led to a constant growth in the amount of computer readable text produced, stored and handled; despite the fact that text information is still widely used in printed paper format. Almost all the printed text information that has been produced over the last twenty years, across the world, has been created using “word processing” software of some form. The motivation for wanting text to be summarised vary according to the field of study and the application of interest; however, it is very clear that, in all cases, what is being pursued is the idea of presenting the text content in a coherent but reduced form.

The focus of the work described in this thesis is the summarisation of the free text found in questionnaires using text classification methods. In the context of this thesis the term “free text” refers to “free form” text that is not restricted by (for example) some database schema; thus the text commonly found in literature, newspaper articles and so on. Questionnaires are a useful and common research tool, used in many fields of study, for collecting information from groups of respondents. Questionnaires contain a series of questions that can either be closed-ended or open-ended [72]. In the case of closed-ended questions the respondents choose an answer from a given set of options; the answers are then stored in a tabular format. On the other hand, in the case of open-ended questions the respondents formulate their own answers in their own words by writing text, therefore the answers are stored in a free text format. Although the classic way in which questionnaires are presented to respondents is in written form, nowadays the use of electronic based questionnaires is also common because of the automated processing advantages offered (no need to first transcribe or scan answers into an electronic format).

The individual purpose of a questionnaire will indicate the way it is constructed, the frequency in which it will be applied and the target group to whom it is directed.

The construction of the questionnaire will take into account the nature of the relevant data that needs to be garnered from the respondents; because, as Gillham [35] states: “*good research cannot be built on poorly collected data*”. The analysis of the information gathered using questionnaires has usually been done using statistical techniques. The extraction of “deeper” information and patterns from questionnaire data (both tabular and free text), beyond straightforward statistical analysis, can be achieved using data mining. The term data mining refers to the extraction or discovery of knowledge from large amounts of data [41]. Knowledge in this context means useful or strategic information that was unexpected or not clearly seen at first.

This thesis is focused on the analysis of the free text element frequently included in questionnaires; because, while extracting meaning from the tabular part of questionnaires is relatively straightforward, extracting meaning from the free text part still presents a challenge especially when we consider that, in most cases, the texts are short, unstructured and contain misspelled words, poor grammar, and abbreviations and acronyms related to a specific domain. Unstructured text refers to text that is not organised in a predefined structure. There are a number of alternative ways in which the extraction of useful information from text can be conducted. For example, using Natural Language Processing (NLP) [65] or Information Retrieval (IR) [6] techniques. However, given the inherent characteristics of questionnaire free text data, as mentioned above, these approaches are unlikely to produce an appropriate result, because as Mitchell *et al.* [75] noted: “With unmoderated mark schemes, the system generates a number of marking errors when faced with unexpected but allowable responses, synonyms and phraseology”. Additionally, Soderland [96] stated: “Unfortunately, standard natural language processing (NLP) extraction techniques expect full, grammatical sentences, and perform poorly on the choppy sentences fragments that are often found on web pages”. The approach proposed in this thesis is founded on the concept of text classification.

In this thesis the term “text summarisation” refers to the automated or semi-automated generation of summaries from free text documents using data mining techniques (although it is acknowledged that techniques other than data mining techniques can equally well be applied to achieve the desired text summarisations, however these are not the focus of this thesis). There are various data mining techniques (and by extension machine learning techniques) that may be adopted to achieve the desired text summarisation, the one proposed in this thesis is text classification. Text classification is primarily concerned with the assignation of one or more predefined categories to text documents according to their content [28]. It can be implemented as an automated process involving none or just a small amount of participation from the domain experts/end user. In the case of questionnaire data, if desired, the tabular part of the data can be processed at the same time in order to add more information to the

extracted knowledge.

With respect to the work described in this thesis text summarisation is conceived of as a form of text classification in that the classes assigned to text documents can be viewed as an indication (summarisation) of the main ideas of the original free text but in a coherent and reduced form. Coherent because the class names that are typically used to label text documents tend to represent a synthesis of the topic with which the document is concerned. Reduced because the use of class labels leads to a succinct articulation of the content of the original text. It is acknowledged that a summary of this form is not as complete or as extensive as what some observers might consider to be a summary; but, if we assign multiple class labels to each document (questionnaire free text) then this comes nearer to what might be traditionally viewed as a summary. It is important to note that although they are intended for different forms of application, text summarisation (as envisioned in this thesis) and text classification share a common purpose, namely to derive meaning from free text (either by producing a summary or by assigning labels). Thus this thesis presents a study of the use of text classification methods for text summarisation with respect to the unstructured free text part of questionnaire data. Four classification summarisation techniques are investigated: (i) using standard classification techniques, (ii) using secondary data to generate classifiers to be applied to primary data, (iii) using a semi-automated rule summarisation extraction tool that requires user interaction, and (iv) using a hierarchical text classification approach to generate the desired summaries. These techniques will be extensively described and analysed in their respective chapters later in this thesis.

The remainder of this introductory chapter is organised as follows. Section 1.2 presents the motivations for the work presented in this thesis. The research question and the related issues are described in Section 1.3. Section 1.4 presents a formal definition of the problem addressed by the research presented in this thesis. The research methodology that was followed is presented in Section 1.5. The contributions of the work, with respect to data mining and related areas of research, are presented in Section 1.6. Section 1.7 gives a general overview of the organisation of this thesis. A number of publications by the author are listed in Section 1.8. Finally, a summary of the chapter is presented in Section 1.9.

1.2 Motivation

The main motivation for the research presented in this thesis, as already noted, is the desire to extract meaning from the free text element frequently found in questionnaire data. A number of related motivations are identified:

1. **Volume of Data:** The growing accessibility and use of computers and the Internet by increasing numbers of people has served to facilitate the collection of

on-line questionnaire data, much of it in free text format. To keep up with the increasing amount of questionnaire data there is a corresponding desire to automate the analysis of this data. The automated analysis of the free text element of questionnaire data remains a challenge.

2. **Fast Analysis:** The desire of public and private institutions to speed up the process of gathering and analysing information (e.g. opinion about politicians, customer satisfaction with a certain product, medical condition prevalence, etc.) to improve decision making requires automated processes, so as those proposed in this thesis.
3. **Resource Reduction:** Reducing resources required for the analysis of the free text element in questionnaires, for example through the use of domain experts, is clearly desirable.
4. **Enhance Analysis:** The potential for the application of much more sophisticated analysis techniques (such as opinion and sentiment mining, and brand reputation mining), directed at specific aspects found in questionnaire free text, will require some form of preliminary processing of this text (such as text summarisation).
5. **Integrated Analysis:** Related to (4), the potential for integrating information extracted from the free text part of questionnaires with the tabular element of questionnaires, or additional datasets, would further facilitate enhanced analysis.
6. **No Previous Investigation:** To the best knowledge of the author, the generation of text summaries using text classification techniques has not been widely investigated in the literature, and thus appropriate investigation is warranted.
7. **Diversity of Application:** The observation that text summarisation, as envisioned in this thesis, has potential benefits in the context of a variety of application domains such as medicine, marketing, national security and finance, among others. The main benefits would be: (i) to provide a greater insight to data that will help in decision making and (ii) to help to fulfil the objectives with respect to the application domain. In some cases unexpected trends could be discovered.

More generally, the mechanisms proposed in this thesis will allow for greater use of questionnaire information gathering as it will reduce the analysis cost. The automated extraction of summaries from unstructured questionnaire text will help people to understand and assimilate the main ideas, contained in large questionnaire collections, in a clear and concise way regardless of the purpose or domain of application. The automated extraction of summaries from the free text element of questionnaire data

is of particular importance in lessening the so called “information load” [25] of the information age; thus, for example, speeding up the process of decision making.

1.3 Research question and issues

Given the above, the research question to which this thesis is directed is: *“can relevant information be extracted, in the form of a summary, from the free text element often found in questionnaires using text classification techniques; while at the same time taking into account that such text is usually sparse, unstructured and contains misspelled words, poor grammar, and abbreviations and acronyms related to a specific domain?”*.

There are five related research issues associated with the provision of an answer to this research question:

1. **Inherent characteristics of the free text part of questionnaires:** As already noted above, the free text part of questionnaires typically features: a lack of structure, misspelled words, poor grammar, use of abbreviations and acronyms, and use of negation. The first issue to be addressed was thus how best to overcome this limitation.
2. **Robust classification techniques:** With respect to the research presented in this thesis, robustness refers to the case when the generation of comprehensive summaries using text classification techniques will require recourse to a considerable number of classes (unlike more traditional text classification applications). Thus a large number of classes is to be expected, and therefore any proposed solution to the above research question must be able to satisfactorily deal with this.
3. **Mechanism for generating the desired summaries from class labels:** The generation of summaries from class labels assigned to the free text part of questionnaires will require an effective mechanism whereby class labels can be used for this purpose.
4. **Unbalanced data:** In the context of the anticipated mechanisms for generating text summaries it is expected that in some cases a significant number of examples per class may not be available, in other cases there will be a large number of records with respect to some classes (classes that represent common occurrences). Therefore, the input data can be expected to be unbalanced. Any proposed classification technique for text summary generation must therefore be able to deal with unbalanced data.
5. **Sufficient training data:** Given that a large number of classes can be expected, sufficient training data must be available so that there are enough examples covering each class. In cases where sufficient training data is not available it will be

necessary to identify questionnaire summarisation techniques that can operate with small numbers of records (examples).

Overall the work described in this thesis can be said to be directed at two principal objectives as follows:

1. The extraction of relevant information from the free text element of questionnaires.
2. The realization of mechanisms to achieve text summarisation using text classification methods.

1.4 Problem definition

In general the techniques presented in this thesis are directed at the summarisation of free text from questionnaire returns, although its applicability can be extended to other types of free text. In the context of Questionnaire Data Mining the input is a collection of n questionnaires, $Q = \{q_1, q_2, \dots, q_n\}$, where each questionnaire comprises a tabular component and a free text component, $q_i = \{Table_i, Text_i\}$ (where i is a numeric questionnaire identifier). The tabular component, in turn, comprises a subset of a global set of m attribute-value pairs $A = \{a_1, a_2, \dots, a_m\}$; thus $Table_i \subset A$. The objective is then to summarise the free text element of the questionnaires in terms of a sequence of p labels (classes), $\{c_1, c_2, \dots, c_p\}$, using a set of k rules, $R = \{r_1, r_2, \dots, r_k\}$, to prepend or append domain-specific text to the class labels.

1.5 Research Methodology

To provide an answer to the research question and associated research issues, as described in the previous sections, the adopted research methodology was to investigate and evaluate a series of techniques directed at the summarisation of free text through text classification. Four techniques in total were considered. The start point for the study was simply to apply straightforward, well established, standard classification techniques and use the generated class labels to form a summary. The aim was to establish a benchmark with which alternative proposed techniques could be compared. The second technique was directed at considering the issue where insufficient data is available by considering questionnaire summarisation in terms of classifiers generated using secondary data. It was also conjectured that the only mechanisms whereby the large number of class labels issue, and the nature of questionnaire free text, could be addressed was using some semi-automated rule summarisation extraction tool that required more user intervention than normally associated with supervised learning. The

fourth technique investigated was founded on the concept of hierarchical text classification. The idea being that a hierarchical approach would be a good way of addressing the large number of classes issue.

So that the proposed techniques could be evaluated a number of data sets were used:

- SAVSNET (Small Animal Veterinary Surveillance Network)
 - SAVSNET-840-4-FT
 - SAVSNET-840-4-TD+FT
 - SAVSNET-971-3-FT
 - SAVSNET-971-3-TD+FT
 - SAVSNET-917-4H
- OHSUMED
 - OHSUMED-CA-3187-3H (Cardiovascular Abnormalities)
 - OHSUMED-AD-3393-3H (Animal Diseases)
- Reuters-21578
 - Reuters-21578-LOC-2327-2H (Locations)
 - Reuters-21578-COM-2327-2H (Commodities)

(Details of these data sets will be presented later in this thesis.) In all cases the free text was typically preprocessed as follows: (i) convert to lower case; (ii) remove numbers, symbols and stop words (common words that are not significant for the text classification/summarisation process), (iii) apply stemming (using an implementation of the Porter Stemming algorithm [106]) and (iv) feature selection. The evaluation metrics used were overall Accuracy, Area Under the receiver operating Curve (AUC), Precision, Sensitivity/Recall and Specificity.

1.6 Contributions

The main contribution of the research work presented in this thesis is the use of text classification methods for text summarisation. As noted earlier, there is little reported work on extracting meaning, in the form of summaries, from free text by using text classification methods. More specifically the work described makes the following contributions:

1. An evaluation of the use of the concept of text classification in the context of questionnaire free text summarisation.

2. A demonstration of the benefits of the usage of classification for summarisation that addresses the issues associated with the nature of the free text element of questionnaire data, which tends to be unstructured and includes misspellings, abbreviations and domain specific terminology.
3. An investigation into the use of secondary data to support free text classification and specifically questionnaire free text summarisation.
4. An investigation into the incorporation of domain experts into the text classification summarisation process.
5. A hierarchical classification mechanism to provide text summarisation with respect to questionnaire data.
6. Implementations of a number of approaches to generating summaries from the free text element of questionnaire data, namely: (i) standard classification techniques, (ii) use of secondary data, (iii) semi-automated summary generation (requires end user involvement) and (iv) hierarchical classification for text summarisation.
7. An investigation into the SAVSNET [83] questionnaire data collections.
8. Supporting investigations using the OHSUMED and Reuters data collections.

1.7 Organisation of the thesis

The rest of this thesis is organised as follows. Chapter 2 presents a literature review of the relevant areas that are of concern with respect this thesis (Questionnaire Data Mining, Text Classification and Text Summarisation). Chapter 3 describes the pre-processing of questionnaire data and a description of the data sets used. Chapter 4 presents the first text summarisation technique proposed, which made use of standard classification techniques, the performance of which provided a benchmark with which to compare the following techniques. Chapter 5 describes the text questionnaire summarisation technique which considers Classifier Generation Using Secondary Data (CGUSD). Chapter 6 then describes the third text summarisation technique that uses a Semi-Automated Rule Summarisation Extraction Tool (SARSET). Chapter 7 presents the fourth proposed technique that uses hierarchical text classification for text summarisation. Experiments and results for each proposed technique, in comparison with the benchmark technique, are included in Chapters 5, 6 and 7 respectively. Finally, in Chapter 8, a comparison of the proposed techniques is presented along with some conclusions and future work.

1.8 Publications

A number of papers and technical reports were produced as part of the research described in this thesis:

1. **M. F. Garcia-Constantino. Technical report of the text summarisation of the free text section of questionnaires related to Veterinary practice. Department of Computer Science, University of Liverpool. Technical report, 2010.** In this technical report a number of text summarisation approaches were reviewed taking into account their possible application in summarising the free text of questionnaires related to a specific domain area, in this case Veterinary Science. At the end of the report the most likely effective text summarisation techniques, to be used in this context, are presented. A significant portion of the material in this report is utilised in Chapter 2 of this thesis.
2. **M. F. Garcia-Constantino, F. Coenen, P. Noble, A. Radford, C. Setzkorn, and A. Tierney. An Investigation Concerning the Generation of Text Summarisation Classifiers using Secondary Data. Proceedings. Machine Learning and Data Mining in Pattern Recognition, Lecture Notes in Computer Science 6871 Springer 2011, pp387-398.** This paper addressed the issue where it was desired to summarise the free text element of questionnaires, but no suitable training data was available. Text summarisation classifiers are generated using secondary data and then applied to the primary data for the purpose of text summarisation. The primary data source was real questionnaire data. The secondary data was extracted from an online bibliographic database of biomedical information. Material presented in this paper has been incorporated in Chapter 5 of this thesis.
3. **S. Chua, F. Coenen, G. Malcolm, and M. F. Garcia-Constantino. Using Negation and Phrases in Inducing Rules for Text Classification. Proceedings. AI-2011: Research and Development in Intelligent Systems XXVIII (Incorporating Applications and Innovations in Intelligent Systems XIX), Springer 2011, pp153-166.** Although not specifically directed at questionnaire text summarisation, this paper investigated the use of negated features in the Inductive Rule Learning (IRL) process by dynamically identifying the features to be negated. The paper also considered text classification based on a “bag of phrases” representation motivated by the observation that a phrase contains more information than a single keyword. The significance with respect to this thesis is that the concept of negation and the bag of phrases representation in text classification has application with respect to text summarisation. Material presented in this paper has been incorporated into the previous

work chapter of this thesis (Chapter 2).

4. **M. F. Garcia-Constantino and F. Coenen. A Survey on Questionnaire Data Mining. Department of Computer Science, University of Liverpool. Technical report, 2011.** This technical report gives a general background, literature review and a categorisation of questionnaire data mining techniques. It also discussed the relevant data mining techniques based on the part of the questionnaire to be mined (tabular, free text or both). A significant portion of the material in this report is utilised in Chapter 2 of this thesis.
5. **M. F. Garcia-Constantino, F. Coenen, P. Noble, A. Radford, and C. Setzkorn. A Semi-Automated Approach to Building Text Summarisation Classifiers. Proceedings. Machine Learning and Data Mining in Pattern Recognition. Lecture Notes in Computer Science. Springer 2012.** In this paper a semi-automated summarisation extraction technique to generate text summarisation classifiers was presented. A realisation of this technique, SARSET (Semi-Automated Rule Summarisation Extraction Tool), was also presented and evaluated using real questionnaire data. The motivation for the work was the observation that the generation of rule-based classifiers from the SAVSNET data might benefit from the intervention of domain experts. Material presented in this paper has been incorporated in Chapter 6 of this thesis.
6. **M. F. Garcia-Constantino, F. Coenen, P. Noble, A. Radford, and C. Setzkorn. A Semi-Automated Approach to Building Text Summarisation Classifiers. Journal of Theoretical and Applied Computer Science. Vol. 6, No. 4, pp. 7-23, 2012.** This paper is a continuation and extension of the previous paper about the SARSET technique (paper number 5 in this list). Additional data sets and classification methods were used for the purpose of evaluation and comparison with respect to SARSET. Material presented in this paper has been incorporated in Chapter 6 of this thesis.
7. **M. F. Garcia-Constantino, F. Coenen, P. Noble and A. Radford. Questionnaire Free Text Summarisation Using Hierarchical Text Classification. Proceedings. AI-2012: Research and Development in Intelligent Systems XXIX (Incorporating Applications and Innovations in Intelligent Systems XX), Springer 2012, pp35-48.** This paper presented an approach to the generation of text summaries using a hierarchical text classification approach. Compared to other text summarisation techniques, the proposed approach provided a more intuitive way of understanding free text and a greater ability to handle large document sets because of the inherent support for the “divide-and-conquer” strategy provided by hierarchical classification techniques. Material presented in this paper has been incorporated in Chapter 7 of this thesis.

8. **M. F. Garcia-Constantino, F. Coenen, P. Noble and A. Radford. Free Text Summarisation of Structured and Unstructured Free Text Using Hierarchical Classification. In preparation.** This unpublished paper presents an investigation into the summarisation of structured and unstructured free text using hierarchical text classification. This paper extends the technique presented in the previous paper concerning the hierarchical classification technique used for text summarisation. Apart from SMO, three other classification techniques are also considered, namely: (i) C4.5, (ii) Naive Bayes and (iii) RIPPER. Material presented in this paper has been incorporated in Chapter 7 of this thesis.

1.9 Summary

This chapter has presented a general overview and a background to the research described in this thesis. The motivation for the work and the research question to be addressed were presented, together with a concise explanation of the adopted research methodology and the contribution of the research. The ideas introduced in this chapter are covered extensively in the remainder of this thesis, commencing with a review of existing work in the following chapter.

Chapter 2

Literature Review

2.1 Introduction

This chapter presents a review of the relevant background knowledge with respect to the research addressed in this thesis. Section 2.2 presents the concept of Data Mining and the KDD (Knowledge Discovery in Databases) process in order to give an overview of the foundations of the research and of the state of the art with respect to techniques typically used to extract useful information from data. Section 2.3 describes the extraction of useful information from questionnaires using data mining techniques, places the Questionnaire Data Mining research domain within the data mining landscape and introduces the idea of mining tabular and textual information from questionnaires. In Section 2.4 feature selection techniques for text classification are reviewed as well as the most relevant approaches and evaluation measures. Section 2.5 describes existing work on text summarisation and the evaluation measures used, and it also provides a review of a number of text summarisation approaches that use text classification methods but in a different manner to that described in this thesis. Finally, a summary of the chapter is presented in Section 2.6.

2.2 Data Mining and the Knowledge Discovery in Databases (KDD) process

Data mining refers to the extraction or discovery of knowledge from large amounts of data [41]. Knowledge in this context means useful or strategic information, that was unexpected or not clearly observable in the first instance, which can be used for decision making. KDD is defined by Fayyad *et al.* [27] as the “*non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*”, where data mining is the step in which the algorithms for pattern extraction are applied to data in order to produce the patterns. According to Han *et al.* [41], the KDD process comprises the following steps: (i) data cleaning, (ii) data integration, (iii) data selection, (iv) data transformation, (v) data mining, (vi) pattern evaluation and

(vii) knowledge presentation. It is beyond the scope of this thesis to describe each step of the KDD process, thus, the rest of this section concentrates on the data mining step within the overall KDD process only.

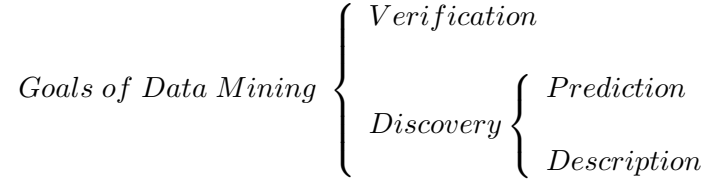


Figure 2.1: Goals of Data Mining

The goals of data mining, according to Fayyad *et al.* [27], are shown in Figure 2.1. In the case of verification, a certain hypothesis is required to be validated by the system; whereas, discovery is directed at the identification of patterns. Discovery is divided into prediction and description; while in the former the patterns discovered by the system are used to predict future outputs, in the latter they are used to provide human understandable descriptions. The data mining goals pursued by the research described in this thesis can be said to be directed at both prediction and description.

Despite the fact that many researchers [27, 41, 85] identify different data mining areas of concern (classification, regression, clustering, association, summarisation, dependency modelling), three of these can be identified as the most dominant:

- **Classification:** The supervised process of finding a model (classifier) based on labeled or pre-classified “training data” in order to be able to predict the classes to which unlabelled data belongs [41].
- **Association:** The unsupervised process of finding frequently occurring itemsets in data so as to produce what are called “association rules” of the form $A \Rightarrow B$, where A and B are non-overlapping itemsets. To measure how interesting or relevant a rule is, support and confidence measures are typically used [41].
- **Clustering:** The unsupervised process whereby data is grouped into clusters based on the similarities between the items that comprise the data. The similarities are typically defined according to some distance function (such as Euclidean distance) [41].

As stated in Chapter 1, the approaches presented in this thesis utilise text classification methods for the purpose of text summarisation. Therefore, the data mining technique which is of greatest significance with respect to this thesis is classification.

Data mining techniques can also be categorised according to the kind of data and/or application domain at which they are directed. Sometimes a domain can be considered

as a part of another domain because it is addressing a specific issue (for example, opinion mining can be argued to be akin to text mining); furthermore, a domain can be the result of combining two domains (for example, spatio-temporal data mining is derived from spatial data mining and temporal data mining). The most significant data mining domains within the overall context of data mining can be considered to be: spatial, text, temporal, spatio-temporal, audio, video, multimedia (audio and video), opinion, data stream, graph and web mining. The data mining domain of interest with respect to the research described in this thesis is text mining.

According to Han *et al.* [41], text mining is concerned with the discovery of “*knowledge from semistructured text data using methods such as keyword-based association analysis, document classification, and document clustering*”. Similar to this description are the definitions of Roiger and Geatz [85], and Hotho *et al.* [52], which can be summarised as defining text mining as “*the extraction of useful patterns from free text by using algorithms and methods from the machine learning and statistics fields*”. Related areas that have contributed to the improvement of text mining approaches are: Natural Language Processing (NLP), Information Retrieval (IR) and Information Extraction (IE).

Natural Language Processing (NLP) is defined by Joshi [65] as: “*the study of mathematical and computational modelling of various aspects of language and the development of a wide range of systems*”. Hotho *et al.* [52] define Information Retrieval (IR) as: “*the finding of documents which contain answers to questions and not the finding of answers itself*”. Hotho *et al.* [52] also state that the goal of Information Extraction (IE) is: “*extraction of specific information from text documents*”.

Text mining is a broad area of research filled with different approaches and techniques to address the challenges imposed by particular problems. In the context of the research described in this thesis, the classification and summarisation of the textual element of questionnaires are of interest. Thus the research presented in this thesis can also be considered to fall within the areas of *Questionnaire Analysis* and *Questionnaire Data Mining* when applied to questionnaires that contain free text (questionnaires can of course be comprised entirely of closed-ended questions). In Section 2.3 the relationship between Questionnaire Data Mining and Text Mining is considered in more detail. Text classification and text summarisation are considered in Sections 2.4 and 2.5 respectively.

2.3 Questionnaire Data Mining (QDM)

As noted in Chapter 1, questionnaires are a useful and common research tool for collecting information from a group of respondents. Questionnaires typically consist of closed-ended and open-ended questions, answers are stored in tabular and free text formats respectively. Questionnaires are used in many fields so as to gather information

from a target group of respondents. The analysis of responses is typically conducted in order to aid and improve some decision making process.

Many different data types can be identified within both the tabular and the free text element of questionnaires. Chen and Weng [14] note that questionnaire tabular data typically includes more data types than those that are usually encountered in data mining (nominal/boolean, ordinal, quantitative). Based on Marshall's [72] categorisation of possible types of tabular data that can result from a questionnaire, Chen and Weng [14] defined seven possible types of tabular data: (i) nominal (categories), (ii) multiple-choice (lists), (iii) quantitative (numbers), (iv) ordinal (ranks), (v) fuzzy ordinal (linguistic ranks), (vi) multiple-choice ordinal (multiple-ranks) and (vii) multiple-choice fuzzy ordinal (multiple-linguistic ranks). On the other hand, while it is clear that the free text part of a questionnaire contains text, in many cases there are numerical or other types of data mixed within the text that might, during the pre-processing, either be considered to be noise and thus removed or considered relevant and be identified ready for further analysis. The pre-processing of the tabular and free text part of questionnaires is extensively covered in Chapter 3.

An interesting trend in questionnaire data mining research is that most of the proposed techniques have been developed in Japan. A possible reason is the popularity of the Kansei Engineering method [78] in Japan. This method aims at the design and production of products based on the feelings and impressions of consumers. The favoured mechanisms for gathering consumer feedback is through questionnaires and surveys, either written or electronically-based, to which statistical and, more recently, data mining techniques have been applied to extract useful information.

While the extraction of useful information from the tabular part of questionnaires is straightforward (for example using well established data mining techniques), extracting useful information from the free text part is more challenging because, as already noted, this kind of text is usually sparse, unstructured and typically contains misspelled words, poor grammar, and abbreviations and acronyms related to a specific domain.

2.3.1 Overview of QDM approaches

There is considerable reported work on extracting meaning from questionnaire data focusing either: (i) only on the tabular part [14], (ii) only on the free text part [1, 34, 32, 30, 45, 50, 104, 108] or (iii) on both [49, 46, 86, 100]. Taking as a foundation the questionnaire data mining categorisation presented by Chen and Weng [14], Figure 2.2 presents the most relevant approaches according to the part of the questionnaire that they are focused on. Thus, the main categories are based on the type of questions: closed-ended (tabular data), open-ended (free text) or both. As noted above, most of the proposed approaches found in the literature are from Japanese researchers and many have been published in Japanese language research venues [37, 47, 48, 56, 55, 54,

57, 77, 79, 88, 87, 101, 102, 109], the review presented here only considers research that has been published at English language venues. The covered approaches are explained and discussed in more detail in the following sections. The approaches proposed by the author that appear in Figure 2.2 are extensively discussed in Chapters 5, 6 and 7.

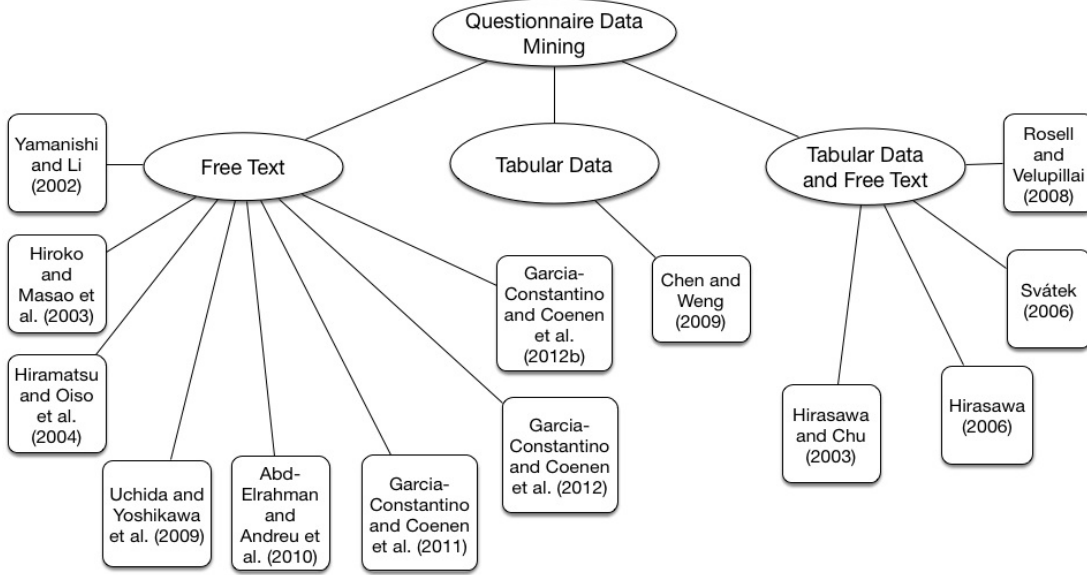


Figure 2.2: Questionnaire Data Mining relevant approaches.

2.3.2 QDM approaches directed at closed-ended questions (tabular data)

As already noted, the application of data mining techniques to the tabular element of questionnaires does not present a particular challenge, tabular data mining is well understood. It can be conjectured that this is the reason for the scarcity of reported work that addresses specifically the mining of the tabular part of questionnaires. However, Chen and Weng [14] proposed the use of Fuzzy Association Rule Mining (FARM) to extract knowledge from tabular questionnaire data. In order to do so they devised the CLL (Category, List and Linguistic) algorithm, which is based on the classic Apriori algorithm [3]. The main claimed distinctions between the CLL algorithm and the Apriori algorithm were: (i) while the Apriori algorithm only takes into account nominal/boolean data, the CLL algorithm considers more data types; (ii) unlike the Apriori algorithm, the CLL algorithm uses similarity functions to measure the similarity between items; and (iii) in the Apriori algorithm an itemset can be contained or not in a transaction, on the other hand, the CLL algorithm can partly include an itemset in a transaction. The results of the performance evaluation of the CLL algorithm were compared to those of the Apriori algorithm, showing a significantly better performance with respect to the CLL algorithm.

2.3.3 QDM approaches directed at open-ended questions (free text)

Mining the answers to open-ended questions (free text) can be done by applying established text analysis approaches, for example, we can attempt to use Natural Language Processing (NLP) [65] or Information Retrieval (IR) [6] techniques. However, given the unstructured nature of questionnaire free text data these approaches are difficult to apply, and thus unlikely to produce an effective result. There are a number of approaches focused specifically on mining the free text part of questionnaires. It should be noted that, with respect to the analysis of the questionnaires designed to support the Kansei Engineering method in Japan, the free text is in Japanese, a language whose word representation and text structure is very different from that of English, thus requiring a very different form of analysis.

A well known automated approach to extract useful information from the free text part of questionnaires was presented by Yamanishi and Li [108], in which two statistical learning techniques (rule analysis and correspondence analysis) were combined and applied to a balanced set of questionnaires where the free text element had been separated into phrases and words to facilitate the extraction of characteristics for individual analysis targets (objects from questionnaires, for example cars) and relationships between the characteristics of the targets. Survey Analyser (SA), a system for survey analysis based on the two statistical learning techniques, was developed to support the approach. The structure of the free text questionnaire answers considered in [108] does not seem to be an issue because the given answers are short (one sentence). The rule analysis technique is intended to discover classification and association rules. The discovered classification rules were considered to be a representation of the characteristics of the target, whilst the discovered association rules were considered to be a representation of the strengths of the associations between the open answers and the target. The correspondence analysis technique consisted of creating a table listing the co-occurrences of data between the targets and the extracted keywords from the free text. The targets and the keywords were then represented in a two-dimensional *positioning map* where the strengths of the association between elements was measured according to how close they were (in a similar way as the distance measurement used in clustering techniques).

The objective of the semi-automated approach presented by Hiroko *et al.* [50] was the extraction of people’s intentions from open-ended questionnaire responses, such as complaints or requests. To achieve this objective, a criterion based on paraphrasing was developed for judging request intentions in response texts. Hiroko *et al.* assumed that responses with request intentions could be paraphrased into expressions used more commonly for requests, such as “I would like to...”. On the other hand, responses that do not have an implicit request intention could not be paraphrased in this way. Three aspects were taken into account to evaluate the criterion: (i) objectivity, (ii) reproducibility and (iii) effectiveness. Objectivity refers to how machine learning methods

can learn the criterion from text that has previously been tagged as indicating an intention. Reproducibility refers to the consistency of the judgement evaluation carried out by human subjects. Effectiveness refers to the ability of achieving the same results without using the criterion. The machine learning methods used were the Maximum Entropy Method (MEM) and Support Vector Machines (SVMs).

Hiramatsu *et al.* [45] presented a system to support the analysis of open-ended questions by extracting only atypical or unexpected opinions present in the answers. The system classified opinions as *typical* or *atypical*. Hiramatsu *et al.* emphasized the significance of open-ended questions by noting that they reflect directly the opinions of the respondents (unlike closed-ended questions). Typical opinions were defined in terms of: (i) the similarity of the answers in closed-ended and in open-ended questions, (ii) the existence of such opinions in an “opinion base” holding all the opinions given in previous applications of the questionnaire and (iii) the relevance of opinions regarding the domain of the questionnaires. Atypical opinions were simply defined as being opinions not considered as typical. Three methods to extract atypical opinions were presented: (i) according to the ratio of typical word combinations in the sentences making up an answer (the basic method), (ii) based on the keyword distance obtained after identifying keywords in the opinions and comparing them with words contained in a “typical word” database and (iii) based on the use of delimiters to split sentences containing opinions into phrases. From the results of the experiments carried out to compare the three methods, Hiramatsu *et al.* found that the third method provided the most effective result.

The approach proposed by Uchida *et al.* [104], based on co-occurrence analysis, is a semi-automated system for extracting the keywords from the free text element of questionnaires and visualising the relationship among sentences. A text mining technique called the Hierarchical Keyword Graph (HK Graph) technique was used to extract the keywords and to represent them in a hierarchical structure. In the HK Graph technique the free text was first divided into words, keywords were then identified by human users. Next the co-occurrence between the selected keywords and other words in the text was calculated; words with the highest co-occurrence values were extracted and represented as a hierarchical graph structure. A set of statistical techniques, known as Multi Dimensional Scaling (MDS), was used to interactively cluster the respondents (people who answered the questionnaires) in a visual space according to the similarity between the extracted keywords, in order to visualise the relationship among the extracted keywords from each cluster, and therefore the tendencies of the opinions given by the respondents.

Another approach based on co-occurrence analysis was presented in [1], where Abd-Elrahman *et al.* compared domain expert’s interpretations of free text to the automated performance of a keyword co-occurrence text mining algorithm implementation

included in the Wordstat software¹. The questionnaires were pre-processed by organising the text and by correcting misspelled words. The free text interpretation included the manual identification of categories and subcategories relevant to the domain to increase the resolution of the analysis. A “keywords” list and an “excluded words” list were created to identify words relevant to the domain as opposed to significant words defined in terms of their linguistic value. A co-occurrence text mining algorithm was then applied automatically.

2.3.4 QDM approaches directed at both closed-ended and open-ended questions (tabular data and free text)

Four approaches addressing questionnaire mining by combining tabular data and free text can be identified from within the literature [46, 47, 86, 100]. In [47], Hirasawa and Chu presented an automated method based on Probabilistic Latent Semantic Indexing (PLSI) [51] to extract useful information from documents with both fixed (tabular) and free (free text) formats, such as questionnaires, by representing both the tabular data and the free text as matrices, merging them, weighting their contents and clustering them according to similarity measures. The clusters were then analysed using statistical techniques and knowledge was extracted from them. PLSI is based on the Latent Semantic Indexing (LSI) model in which each document and query vector are mapped in order to reduce the dimension of the vector space by using Single Value Decomposition (SVD). Hirasawa and Chu noted that the performance of their method was dependent on the structure of the documents subject to the analysis.

In [46] Hirasawa presented two algorithms: one for processing both tabular data and free text, and another for extracting important sentences from questionnaire data. The algorithms were combined with statistical techniques, thus forming an automated method for extracting useful information from questionnaires. In a similar way as in [47], PLSI was used not only to cluster the questionnaires, but also to classify them if desired. Important sentences were extracted from the free text using one of the proposed algorithms, and statistical techniques such as multiple linear regression were applied to the tabular data. The outputs of the clustering and classification processes were combined separately with the outputs of the extracted sentences from the free text and with the outputs of the statistical techniques applied to the tabular data to extract useful information.

In the approach presented by Svátek [100], the objective was to use both the tabular data and the free text part of questionnaires to build a domain ontology to discover association rules in questionnaire tabular data. An ontology is a set of concepts and their relationships within a domain. Svátek stated that the text found in both the tabular and free text part of questionnaires can be used as ontology components (e.g. classes,

¹<http://www.kovcomp.co.uk/wordstat/wordbroc.html>.

relations and instances). In this context, the ontologies were used as a mechanism for both: (i) semi-automatically focusing the mining process and (ii) for assisting in the interpretation of the discovered knowledge. Two case studies were presented where the ontology design process was conducted manually, it is suggested that a tool based on POS (Part-Of-Speech) tagging might be implemented to support the technique.

The fourth identified dual questionnaire mining approach was that of Rosell and Velupillai [86], who proposed a semi-automated system aimed at the generation of hypothesis from questionnaire data by applying text clustering to the free text element and classification to the tabular element. While the text clustering process was semi-automated, the evaluation of the clusters generated against the classified tabular data was automated. The method comprised four steps: (i) clustering the free text, (ii) identifying interesting clusters, (iii) exploring the content of the clusters and (iv) formulating hypotheses. The free text was represented as a text-by-word matrix using the vector space model. An algorithm, the Relative Clustering Algorithm, was used to construct a cluster of the words related to a text cluster in order to be used as a description of the cluster. Tabular data was used as a way of categorising the clusters generated for the free text.

2.4 Text Classification

Text classification can also be referred to as text categorisation; however, in this thesis the term text classification is used throughout. According to Fragoudis *et al.* [28], text classification is “*the task of assigning one or more predefined categories to natural language text documents, based on their contents*”. The survey presented by Sebastiani [90] indicated that using machine learning techniques for automated text classification has more advantages than the approaches that rely on domain experts to manually define a classifier. In this survey the problems that are discussed are: document representation, classifier construction and classifier evaluation. Similar to the classification applied within the context of the mining of tabular data, the approaches in text classification can be divided according to whether the texts have a single label or multiple labels associated with them as shown in Figure 2.3.

$$\text{Text Classification} \left\{ \begin{array}{l} \text{Single - labelled} \left\{ \begin{array}{l} \text{Binary classification} \\ \text{Multi - class classification} \end{array} \right. \\ \text{Multi - labelled} \end{array} \right.$$

Figure 2.3: Text Classification.

In single-labelled text classification a document can only be related to one specific label. In *binary text classification* each document is assigned to either a specific pre-defined category (single label) or to the complement of that category [90]. As noted by Wang [105], this type of text classification can be considered as a two-class (positive or negative) approach. On the other hand, multi-class classification refers to the situation where each document is assigned a category from a set of n classes (where $n > 2$). Multi-labelled text classification refers to the case in which a document can be associated with more than one label. In the subsections below the following topics related to text classification are considered: feature selection, algorithms, evaluation measures and approaches.

2.4.1 Feature Selection in Text Classification

Feature selection refers to the process of selecting relevant features from text where typically each term (word/phrase) in the text represents a feature. The aims are to improve both the effectiveness of the classification and the efficiency in computational terms (by reducing the dimensionality) [84]. Miner *et al.* [74] categorised text mining feature selection approaches according to whether they were based on:

1. **Information theory:** Addressing the best way to process signals and compress and communicate data.
2. **Statistics:** Determining the statistical correlation between the terms and the class labels of the documents.
3. **Frequency:** Determining the importance of the terms based on their frequency and on the document frequency.

In relation to the Information Theory and the Statistics methods, Frequency methods are less computationally expensive. The most relevant approaches with respect to these categories and the work described in this thesis are described in the following subsections, namely: (i) *Information Gain (IG)*, (ii) *Chi-squared (χ^2)*, (iii) *Correlation-based Feature Selection (CFS)* and (iv) *Term Frequency-Inverse Document Frequency (TF-IDF)*. A justification of which feature selection methods were used for the work described in this thesis, and how they were used, is presented at the end of this section.

2.4.1.1 Information Gain (IG)

Information Gain (IG) is based on Information Theory, which is concerned with the processing and compression of signal and communication data, and was introduced in 1948 by Claude Shannon [91] who is considered to be the “father” of information theory. According to Miner *et al.* [74] IG “measures how much the uncertainty about the target variable, called *entropy*, is reduced when the feature is used”. In other words, given

that entropy is a measure of uncertainty with respect to a training set (or the amount of information required to assign a class label to an instance), IG is an indicator of how much information is gained from an initial to a new entropy of a feature. It is calculated as follows:

$$Gain(A) = Info(D) - Info_A(D) \quad (2.1)$$

where D is a data partition which comprises instances in a node N , which represents tuples of partition D . The information required to assign a class label to an instance in D , in other words the entropy of D , is given by:

$$Info(D) = - \sum_{i=1}^m p_i \log_2(p_i) \quad (2.2)$$

where a class label can have m different values and p_i is the probability that an instance belonging to D is related to a certain class. The information required to produce a correct classification is given by:

$$Info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times Info(D_j) \quad (2.3)$$

$\frac{|D_j|}{|D|}$ represents the weight of the j th partition. A feature with high IG has a better occurrence prediction of the target variable. Typically, features are ranked according to their IG and the features with higher values (which have a better prediction capability with respect to the class labels) are chosen.

2.4.1.2 Chi-squared (χ^2) statistic

The Chi-squared (χ^2) statistic measures the lack of independence between a feature and a class to which a document is related [111]. The more independent, the more irrelevant a feature is with respect to a certain class. The Chi-squared statistic is calculated for each term and, after ranking all the features, the most relevant are chosen. In the formal definition of Chi-squared, two features A and B are considered; they can have different values and are paired: (A_i, B_j) , where A and B can take any value a or b respectively, from 1 to c in the former and from 1 to r in the later. As explained in [41], the Chi-squared statistic is then calculated as:

$$\chi^2 = \sum_{i=1}^c \sum_{j=1}^r \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \quad (2.4)$$

where, with respect to (A_i, B_j) , o_{ij} is the *observed frequency* and e_{ij} is the *expected frequency*. e_{ij} is calculated as:

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N} \quad (2.5)$$

where N is the number of instances, $count(A = a_i)$ is the number of instances where the value for A is a_i and $count(B = b_j)$ is the number of instances where the value for B is b_j .

In the comparative study of feature selection methods presented by Yang and Pedersen [111] the performance of the Chi-squared statistic is similar to IG when used as a ranking metric. Miner *et al.* [74] points out the correlation of the computational cost of the Chi-squared statistic with the size of the vocabulary.

2.4.1.3 Correlation Feature Selection (CFS)

Correlation Feature Selection (CFS) is used to identify and select sets of features which are “highly correlated with the class but with low intercorrelation” [107] in order to remove redundant or irrelevant features. Redundancy in this context is given by a feature being highly correlated with one or more features. As presented by Witten *et al.* in [107], considering two nominal attributes A and B , their correlation is measured using *symmetric uncertainty*, which is defined as:

$$U(A, B) = 2 \frac{H(A) + H(B) - H(A, B)}{H(A) + H(B)} \quad (2.6)$$

where H represents the entropy function and $H(A, B)$ the joint entropy of A and B . Symmetric uncertainty can take values between 0 and 1. The relevance of a set of features using CFS is determined by:

$$CFS = \frac{\sum_{j=1}^m U(A_j, C)}{\sqrt{\sum_{i=1}^m \sum_{j=1}^m U(A_i, A_j)}} \quad (2.7)$$

where the C in the numerator indicates the class and the (A_i, A_j) indicates a pair of attributes in the set of features. If in a selected set of features there is a correlation between all the m attributes and the class, the numerator (the total symmetric uncertainty) is then m and the denominator $\sqrt{m^2}$, thus the CFS value will be 1, which is the maximum symmetric uncertainty value that can be obtained. In other words it is not possible to distinguish between classes. It is therefore better to focus on smaller subsets of features in order to find subsets with low symmetric uncertainty that are highly correlated with a class label but have a low correlation between them.

2.4.1.4 Feature Weighting with Term Frequency-Inverse Document Frequency (TF-IDF)

The Term Frequency-Inverse Document Frequency (TF-IDF) statistic weights terms by combining how frequent a term is in a document (TF) with how rare the term is with respect to the entire document set (IDF) [8]. TF-IDF is calculated as:

$$TF - IDF(d, t) = TF(d, t) \times IDF(t) \quad (2.8)$$

where d represents a document, t represents a term, TF is the term frequency and IDF is the inverse document frequency. Term Frequency (TF) is the number of occurrences of a term (feature) in a document and is calculated as:

$$TF(d, t) = \sum_{i \in d}^{|d|} 1\{d_i = t\} \quad (2.9)$$

Document Frequency (DF) is the number of documents that contain a particular term. Inverse Document Frequency (IDF) [63], on the other hand, address the issue of DF not being a good discriminator by considering the importance of terms in relation to the total number of documents and to the number of documents in which the term is contained. IDF is calculated as:

$$IDF(t) = \log \frac{1 + |d|}{|d_t|} \quad (2.10)$$

where d is the total number of documents and d_t is the number of documents in which the term t is contained. The resulting TF-IDF weight is assigned to each unique term in the document set and all the terms are ranked from the highest to the lowest weight value indicating their relevance. A user defined threshold k is used to select the top k terms.

2.4.1.5 Feature selection methods used

For comparison purposes with respect to the summarisation techniques proposed in this thesis two alternative feature selection techniques were considered: (i) Term Frequency-Inverse Document Frequency (TF-IDF) [63] plus Chi-squared [113] and (ii) TF-IDF plus Correlation-based Feature Selection (CFS) [40]. Both combine TF-IDF with another feature selection technique, namely Chi-squared and CFS. Chi-squared was chosen because it is an established and widely used feature selection method that calculates the Chi-squared statistic of each feature in relation to a given class in order to identify the features that have relevance with respect to the class. CFS, on the other hand, identifies subsets of uncorrelated features amongst each other that, as a subset, are highly correlated with a class. CFS is not as widely used as Chi-squared but presents an interesting and different idea with respect to the selection of relevant features. Information Gain was not used because experiments (not reported in this thesis) were found to produce a very similar performance to that obtained using Chi-squared, thus corroborating the study by Yang and Pedersen [111].

The reasons for using two different feature selection techniques in conjunction with TF-IDF were: (i) to see how well they performed in conjunction, (ii) to demonstrate

the effectiveness of combining TF-IDF with other feature selection methods to improve the selection of relevant attributes and (iii) to compare how well they performed with data sets containing different types of data. Due to the different nature of the feature selection techniques used in conjunction with TF-IDF, different search methods were used in each case: (i) a ranking search method in the case of Chi-squared and (ii) a genetic search method in the case of CFS.

2.4.2 Relevant Text Classification algorithms

There is no single “best” classification algorithm that can be applied effectively to every data mining problem. The reasons for this are unclear but it is conjectured that this is due to the unique characteristics of individual datasets: (i) the type of data, (ii) the size of the data and (iii) the number and distribution of the classes amongst the records. Research conducted within the data mining community, over the last few decades, has resulted in many different techniques that might suit specific conditions (such as small data sets, data sets comprised mostly of numeric records and data sets that feature a large number of classes). With respect to the work described in this thesis a number of different classification techniques were considered: (i) Bayesian Classifiers (Naïve Bayes), (ii) Decision Trees (C4.5), (iii) Rule Learners (TFPC and RIPPER), (iv) k-nearest neighbour (KNN), and (v) Support Vector Machines (SMO and LibSVM). Each is discussed in more detail in the following subsections. The seven indicated algorithms (Naïve Bayes, C4.5, TFPC, RIPPER, KNN, SMO and LibSVM) were selected for a variety of reasons, as will become apparent in the following subsections.

2.4.2.1 Bayesian Classifiers

Bayesian classifiers are probabilistic classifiers based on Bayes’ theorem, which was proposed and named after Thomas Bayes. Bayes’ theorem is usually expressed as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2.11)$$

where A is a hypothesis, B is evidence, $P(A|B)$ is the posterior probability of A conditioned on B , $P(A)$ and $P(B)$ are the prior probabilities of A and B respectively, and $P(B|A)$ is the posterior probability of B conditioned on A .

Of the many implementations of Bayes’ theorem that have been proposed, the simplest one (Naive Bayes) is used in this thesis mainly to serve as a benchmark with which to compare the other classification techniques. Naive Bayes combines prior and conditional probabilities to calculate the probability of alternative classifications [8]. It is called “naive” because it “naively” makes the assumption that attributes are independent of each other, thus probabilities can be multiplied. Despite the “naive” assumption, the Naive Bayes algorithm has proved to be an effective form of classifier

generator, especially when feature selection has been performed and only non-redundant (independent) attributes are left.

2.4.2.2 Decision Trees

Decision trees have been used for classification purposes for many years. The main advantage of using decision trees is their simplicity. The decision tree classification process can be easily understood and interpreted, and is straightforward to explain. An additional advantage is that, if desired, rules can be easily generated from decision trees (see below). The decision tree algorithm adopted with respect to the evaluation described in this thesis was C4.5. Proposed by Quinlan [82], C4.5 is the successor to the ID3 (Iterative Dichotomiser 3) algorithm [81] and has established itself as a benchmark algorithm throughout the data mining community.

2.4.2.3 Rule Learners

As described in [41], classification rules have a general form: *if* $\langle ANTECEDENT \rangle$ *then* $\langle CONCLUSION \rangle$, thus a simple *if a then b* or a complex *if a and b then c*. A common mechanism for generating rule based classifiers is to use Classification Association Rule Mining (CARM), another is rule induction. As explained in [107], association rules can be used to predict a class attribute, however CARM usually results in a large number of Classification Association Rules (CARs). Support and confidence thresholds are used in order to limit the number of generated rules by keeping the most relevant rules. Support is an indicator of the coverage of a rule, while confidence is an indicator of the accuracy of a rule. Bramer [8] defines support as “the proportion of right-hand sides predicted by the rule that are correctly predicted” and confidence as “the proportion of the training set correctly predicted by the rule”. In the context of the work described in this thesis the rule-based algorithms considered are the TFPC (Total From Partial Classification) and RIPPER (Repeated Incremental Pruning to Produce Error Reduction) algorithms.

The TFPC algorithm was proposed by Coenen *et al.* [18] and is a CARM algorithm based on the Apriori-TFP (Total From Partial) Association Rule Mining (ARM) algorithm [17]. Apriori-TFP, in turn, was founded on the classic Apriori algorithm [3]. While Apriori-TFP generates association rules, TFPC generates classification association rules. The difference between TFPC and other CARM algorithms is that it does not follow the typical approach of first generating all the rules and then pruning them to generate a classifier; instead, TFPC comprises a single step in which all the rules are generated according to a process of identifying the frequent sets of attributes that can be used to generate CARs.

RIPPER is a CAR mining algorithm proposed by Cohen [19] in which “classes are examined in increasing size and a set of rules for a class is generated using incremental

reduced-error pruning” [107].

2.4.2.4 Nearest Neighbour Techniques

Nearest neighbour classification techniques operate using some form of similarity function to compare new instances with existing instances. The similarity is given by the representation of each instance as a point in an n -dimensional space where an unseen instance is classified according to the nearest classified instances. The distance between the points is usually measured using Euclidean distance, but other metrics can be used (for example the Mahalanobis distance or the Chebyshev distance). The most well known nearest neighbour technique is the K-Nearest-Neighbour (KNN) algorithm. The KNN algorithm classifies unknown instances based on their similarity to their closest training instances in a feature space. Because of its nature, it is more computationally expensive than other methods, however KNN is used for evaluation purposes in this thesis because of its enduring popularity.

2.4.2.5 Support Vector Machines (SVM)

Support Vector Machines (SVMs) are a relatively recent addition to the range of available classification techniques compared to other classification techniques. However, SVMs have proved to be very effective in the context of text classification [112]. The SVM technique operates by separating the training instances in an instance space of a binary classification problem using a maximum-margin hyperplane; the hyperplane (among many other existing hyperplanes that can also be used to separate the training instances) that corresponds to the maximum separation between the training instances of the two classes. In mathematics, a hyperplane is typically defined as an $(n-1)$ -dimensional subspace of an n -dimensional vector space. The hyperplanes are based on the instances of both classes that are near the boundaries that separate them. In the context of this thesis two SVM algorithms were used: (i) SMO (Sequential Minimal Optimization) and (ii) LibSVM (Library for Support Vector Machines).

SMO (Sequential Minimal Optimization) was proposed by Platt [80] and is similar to other SVM algorithms in that it divides a large QP (Quadratic Programming) problem into smaller QP problems. SMO differs with respect to other SVM algorithms in that it uses the smallest QP problems, as a result it is much more computationally efficient with respect to both cost and time.

Chang and Lin [11] presented LibSVM (Library for Support Vector Machines), which is a library that includes many SVM implementations for multiclass classification, regression and one-class problems and options for different kernels: linear, polynomial, radial-basis and sigmoid. The SVM implementation and the kernel used with respect to the experiments described in this thesis are classification and radial-basis respectively.

2.4.3 Evaluation measures for Text Classification

In order to assess how well a classification technique performs a number of evaluation measures may be used. Five evaluation measures were used for assessing the performance of the classification processes used for summarisation purposes as described later in this thesis, namely: (i) Accuracy, (ii) Area Under the ROC Curve (AUC), (iii) Precision, (iv) Recall/Sensitivity and (v) Specificity. Each is described in more detail in the following subsections along with the concept of a confusion matrix.

2.4.3.1 Confusion Matrix

A confusion matrix is a 2 by 2 table used to record the performance of a binary classification exercise. The actual classes are listed along the Y-axis (the rows), and the predicted classes along the X-axis (the columns). An example of a confusion matrix is given in Table 2.1:

Table 2.1: Example of a binary confusion matrix

Actual class	Predicted class	
	Positive	Negative
Positive	TP (True Positive)	FN (False Negative)
Negative	FP (False Positive)	TN (True Negative)

From the table:

- TP (True Positive): Number of positive instances that were correctly classified as positive.
- TN (True Negative): Number of negative instances that were correctly classified as negative.
- FP (False Positive): Number of negative instances that were incorrectly classified as positive.
- FN (False Negative): Number of positive instances that were incorrectly classified as negative.

As will become apparent below, all the evaluation measures used in this thesis can be calculated from the confusion matrix values resulting from a classification.

2.4.3.2 Accuracy

Accuracy is the simplest and easiest to understand and obtain of the evaluation measures used. The accuracy of a classifier is a percentage given by the number of instances correctly classified. With respect to the confusion matrix, accuracy is calculated by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.12)$$

The accuracy of a classifier is an indicator of the quality of a classifier. Despite being conceptually simple and easy to calculate, the accuracy measure can be misleading because it does not take into account the distribution of the classes (the “class priors”). It is therefore considered to be a good practice to use at least one other evaluation measure besides accuracy in order to provide an overall reliable indication of the performance of a classifier.

2.4.3.3 Precision

Precision is an indicator of the number of retrieved records (documents in the case of text classification) that are relevant to a given query. In [8], precision is defined as the “proportion of instances classified as positive that are really positive”. Precision is also sometimes referred to as the “Positive Predictive Value” (PPV). With respect to the confusion matrix, precision is calculated as follows:

$$Precision = PPV = \frac{TP}{TP + FP} \quad (2.13)$$

2.4.3.4 Recall/Sensitivity

The recall or sensitivity measure indicates the number of relevant documents that are retrieved by a given query. It is a measure of the proportion of actual positives which are correctly classified as positives, in other words the “True Positive Rate” (TPR). With respect to the confusion matrix it is calculated using:

$$Recall/Sensitivity = TPR = \frac{TP}{TP + FN} \quad (2.14)$$

2.4.3.5 Specificity

Specificity measures the proportion of true negatives which are correctly classified as negatives, in other words the “True Negative Rate” (TNR). In relation to the confusion matrix it is calculated by:

$$Specificity = \frac{TN}{FP + TN} = 1 - FPR \text{ (False Positive Rate)} \quad (2.15)$$

2.4.3.6 Area Under the ROC Curve (AUC)

ROC curves were originally used in the field of signal processing. In the context of binary classifier problems, a Receiving Operating Curve (ROC) plots the False Positive Rate (1-Specificity) against the True Positive Rate (Recall/Sensitivity) along the X and Y axis respectively. If a ROC curve lies along the diagonal this indicates a performance

equivalent to guessing, if it is located above the diagonal the performance is better than guessing, below the diagonal the performance of the corresponding classifier is worse than guessing. By calculating the Area Under the ROC Curve (AUC) a measure can be obtained regarding the effectiveness of a given classifier [42]. An AUC value of 0.5 indicates a guess, thus an effective classifier should have an AUC value associated with it that is greater than 0.5. The nearer the AUC value is to 1.0 the better the performance of the corresponding classifier. The advantage of using the AUC measure for determining classifier performance is that it takes into account the class priors.

2.5 Text Summarisation

The final part of this previous work chapter is directed at a “state of the art” review of work directed at text summarisation. The motivations for text summarisation vary according to the field of study and the nature of the application domain of interest. However, it is very clear that in all cases what is being pursued is the extraction of the main ideas of the original text, and the consequent presentation of these ideas to some audience in a coherent and reduced form. Recall that the type of text which is of principal interest in the context of this thesis is text derived from questionnaires. Thus text that is unstructured, and contains misspelled words, poor grammar and abbreviations and acronyms related to a specific domain. As already noted in Chapter 1, it is acknowledged that the text summaries generated using the techniques proposed in this thesis are not necessarily the same as those generated using more traditional approaches. The remainder of this section is organised as follows. Subsection 2.5.1 provides a categorisation of text summarisation approaches. In Subsection 2.5.2 a number of popular text summarisation approaches are reviewed. Subsection 2.5.3 then considers summarisation techniques that use text classification methods (as in the case of the methods proposed in this thesis), while Subsection 2.5.4 considers evaluation measures for text summarisation.

2.5.1 Categorisation of Text Summarisation techniques

Text summarisation has been a domain of research for many years. An early example (c1958) can be found in [70] in the context of literature abstracts. Many text summarisation techniques have been reported in the literature; these have been categorised in many ways according to: (i) the field of study, (ii) factors inherent to the text or (iii) whether they adopt a statistical or a linguistic approach. Three dominant categorisations of text summarisation techniques are those proposed by: (i) Jones *et al.* [64], (ii) Alonso *et al.* [4] and (iii) Afantenos *et al.* [2]. Jones *et al.* [64] proposed a categorisation dependent on: (i) the input that is received, (ii) the purpose of the summarisation and (iii) the output desired. Afantenos *et al.* [2] and Alonso *et al.* [4]

presented their categorisation of text summarisation in their respective surveys based on the one previously produced by Jones *et al.* [64]. The categorisation of Alonso *et al.* [4] is founded on the “traditional phases” of text summarisation: (i) analysis of the input text (input), (ii) transformation of the input text into the form of summary (purpose) and (iii) synthesis of the output to produce the desired summary (output). Many of the text summarisation approaches presented in the literature focus on one of these phases. The categorisation of Afantenos *et al.* [2] suggests categorisation based on a number of factors. These factors are arranged into three groups (according to the categorisation by Jones *et al.* [64]). The groups are: (i) input, (ii) purpose and (iii) output. The details of each of these groups is shown in Figures 2.4, 2.5 and 2.6.

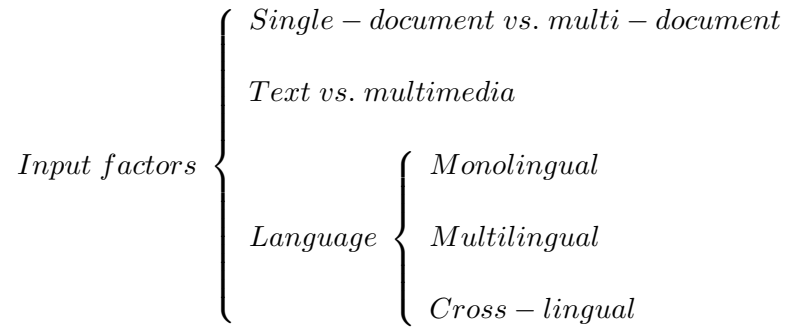


Figure 2.4: Input factors to be considered for text summarisation

With respect to Figure 2.4, the categorisation by single-document or multi-document is self explanatory: the input for summarising can be one document or many documents, respectively. This categorisation can also be found in [71], where it is described briefly, and in [22], where it is used to define the structure of a survey. The text vs. multimedia categorisation is in regard to the format in which the input and the output are presented, whether it is in the form of text or some multimedia format (e.g. image, sound, video). The next categorisation is in regard to the language of the input to be summarised: in the case of monolingual summarisation the language is the same in both the input and the output; in the case of multilingual summarisation the input and output languages are the same, but more than one language may be used; in the case of cross-lingual summarisation the input and the output use different languages.

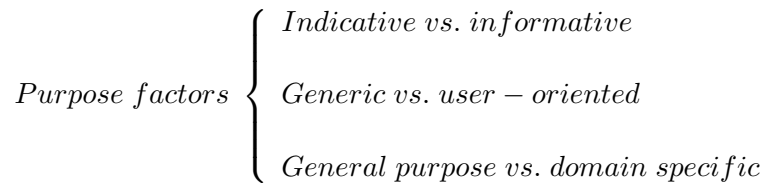


Figure 2.5: Purpose factors to be considered for text summarisation

In the case of the purpose factors (Figure 2.5), a summary can be classified as being indicative if the summary does not replace the original document but indicates the relevant contents; or informative if the source document is replaced but the information covered is taken into account. Generic text summaries are those that give a general view of the text and, as the name suggests, are not intended to fulfil the requirements of a specific type of user. User-oriented summaries, on the contrary, are produced according to the interest of a specific audience. Gong and Liu [36], and Mani [71] also use this categorisation in their approaches. Finally, while a general purpose summarisation system can be used in many domains, a domain specific one can only be applied to a domain of interest.

$$\text{Output factors} \left\{ \begin{array}{l} \text{Completeness} \\ \text{Accuracy} \\ \text{Coherency} \end{array} \right.$$

Figure 2.6: Output factors to be considered for text summarisation

The categorisations within the output group are shown in Figure 2.6 and takes into account the quality of the output, in other words providing a mechanism for measure whether a summary is complete, accurate and coherent (among other things). Mani [71] also addressed the categorisation of text summarisation techniques according to output factors distinguishing between the concept of a summary and an abstract. In their work a summary was formed by extracting the most important sentences and putting them together, typically to the detriment of readability, to form a summary. An abstract, on the other hand, is formed as a result of processing extracted sentences, putting them together and using an algorithm to provide a more human-like interpretation.

In the context of the research described in this thesis the categorisation of Afantenos *et al.* [2] has been adopted. Thus the summarisation techniques described in this thesis can be categorised as follows:

- Input factors
 - Multi-document: because with respect to the research described in this thesis the mining was applied to many questionnaires.
 - Text: because the text summarisation was applied to the free text part of the questionnaires.
 - Monolingual: because the questionnaires used English language only.
- Purpose factors

- Indicative: because the generated summary is an indication of the relevant contents in the free text.
 - Generic: because the text summarisation of the free text did not have the purpose of fulfilling any specific user requirements.
 - General purpose: because the proposed summarisation techniques were intended to be used with respect to questionnaires directed at different fields.
- Output factors
 - Complete: because the generated summaries include the most relevant ideas from the free text.
 - Accurate: because the generated summaries are accurate with respect to the content of the free text.
 - Coherent: because the generated summaries are presented to the user in a readable and understandable way.

2.5.2 Relevant Text Summarisation approaches

A number of popular text summarisation approaches will be presented in this subsection. The aim is to provide an overview of existing approaches with which the approaches proposed later in this thesis can be compared. The different approaches presented here are categorised according to the concepts on which they are based: (i) lexical chains, (ii) sentence extraction and ranking, and (iii) other techniques.

2.5.2.1 Techniques based on lexical chains

The techniques based on the use of lexical chains are closely related to Natural Language Processing (NLP). According to Morris and Hirst [76] a lexical chain is “*a succession of a number of nearby related words spanning a topical unit of the text*”, in other words a lexical chain is a sequence of related words. Many approaches have been proposed that use the concept of lexical chains, which can “*hold a set of semantically related words of a text*” [26]. For example, consider the following text: “*Recently, the Department of Transport has been addressing issues related to aircraft safety. Although the use of airplanes is more common, the use of helicopters in cities has increased in the last couple of years.*”, in this case the lexical chain is “{*transport, aircraft, airplanes, helicopters*}”, where “*airplanes*” and “*helicopters*” are specialisations of “*aircraft*”, which in turn is a specialisation of “*transport*”.

Barzilay and Elhadad [7] proposed an algorithm that identifies lexical chains in a text without the need for full semantic interpretation. Barzilay and Elhadad merged various knowledge sources, the WordNet thesaurus, a part-of-speech tagger, a parser to identify nominal groups and a segmentation algorithm. Their method considers lexical

chains as a source but it ignores any other information from the text. Although their approach was not the first to focus on the use of lexical chains [76], it was very influential with respect to the work of many researchers in the field of text summarisation.

Silber and McCoy [92] proposed a linear time algorithm for calculating lexical chains focusing on its efficiency. Later, in [93], they proposed an improved version of this linear time algorithm and a method for evaluating lexical chains as an intermediate step in summarisation. The algorithm can handle larger documents than that proposed by Barzilay and Elhadad [7], thus, making Silber and McCoy’s approach much more tractable.

Co-reference chains are lexical chains that are related to each other by having the same referent (concept or idea). The approach described in [5] is based on co-reference chains, where a summary representation is constructed by selecting the “best” co-reference chain that best represents the main topic of a text. Thus, the generated summary is a concatenation of sentences from the text that contain one or more elements found on the “best” co-reference chain. The results of the evaluation of the implementation of the algorithm showed that there was a high level of precision with respect to the different criteria taken into consideration.

Another approach related to lexical chains is that presented by Fuentes and Rodríguez [29], in which the cohesive properties of text (namely lexical chains, co-reference chains and named-entity chains) are used to develop a system that allows for the automatic extraction of summaries. Named-entity chains are those that are related to existing categories of names, such as the names of persons, places, etc. The system’s performance was described as positive; nevertheless, further improvement was needed and the fact that it has only been used within the Spanish language context gives it a limited application.

2.5.2.2 Techniques based on sentence extraction and ranking

Many approaches use sentence extraction as a way to fragment the text and then to generate summaries from the extracted sentences. However, some approaches optimise the way that sentences are extracted from the text due to the importance that this process has in the construction and synthesis of the output to produce a summary.

The approach of Jing [59] presented a sentence reduction system that focused on determining the sentences that are less important in the text and that can be removed. In the evaluation of the method, the system’s reduction of a particular text was close to the reduction made by humans. A limitation for using this approach is that it is a generic summarisation approach.

Chuang and Yang [15] focused on sentence extraction from a machine learning perspective and presented the design of an automatic text summariser trained using a supervised learning algorithm. It extracted sentence segments based on a feature vector

representation. The learning algorithms chosen for training the summariser were: the decision tree C4.5 algorithm, the naive Bayesian classifier and the DistAl neural network algorithm. The results of the experiments using the DistAl and the Bayesian learning algorithms outperformed the results that were obtained using the C4.5 algorithm.

Mihalcea [73] proposed an unsupervised method for automatic text summarisation by using a graph-based ranking algorithm for sentence extraction. The sentence extraction algorithm, called TextRank, took into account the local context of a word and the information recursively produced from the entire text. In order to identify important sentences, a process of “recommendation” was used, it consisted of the establishment of a link between two sentences that were related to similar concepts, consequently sentences that have a greater recommendation received a higher score than the ones that will appear in the summary. The TextRank algorithm was portable to other domains, genres or languages because it did not require significant linguistic knowledge.

2.5.2.3 Other techniques

Knight and Marcu [67] suggested a summarisation technique that went beyond sentence extraction; two approaches were proposed for optimizing the process of sentence compression: a noisy-channel based model and a decision-tree based model. In the evaluation of the approaches, the performance of the compression algorithms was closer to human performance. However the performance dropped when there were sentences related to another corpus or set of texts than the ones that were used to train the data: the noisy-channel based model performance was slightly decreased, while the decision-tree based model performance decreased dramatically.

An approach that edits the sentences that result from a sentence extraction process is the one presented by Jing and McKeown [60] in the form of a “cut and paste” based automatic text summariser. The “cut and paste” operations used by the summariser are based on the analysis of abstracts generated by humans: sentence reduction, sentence combination, syntactic transformation, lexical paraphrasing, generalisation or specification (depending on the needs and context), and reordering of the extracted sentences. The text summariser included a decomposition program for the analysis of abstracts written by humans and a sentence reduction module that based its decisions on several knowledge sources. The overall evaluation of the system was satisfactory.

A different approach to text summarisation that involved the use of Singular Value Decomposition (SVD) was taken by Steinberger and Ježek [97]. SVD had been used extensively in the area of statistics. In this approach, two evaluation methods based on SVD were proposed, these methods measured the similarity between the contents of an original document and its summary. The summarising method proposed performed better than the other examined methods considered during testing.

2.5.3 Text Summarisation approaches that use Text Classification methods

To the best knowledge of the author, the generation of text summaries using text classification techniques has not been widely investigated in the literature. However, some examples can be found in [9, 89, 58, 43]. Celikyilmaz and Hakkani-Tür [9] presented an “automated semi-supervised extractive summarisation” approach which used latent concept classification to identify hidden concepts in documents and to combine them to produce summaries. In [89], an approach is proposed composed of a summariser and a classifier integrated within a framework for cleaning and preprocessing data. The point is made that the composition is invertible, meaning that summarisation can be applied first to increase the performance of the classifier or the other way around. In [89] it was also suggested that the use of classification improves the generation of summaries with respect to domain-specific documents. In [43], a system was proposed that identifies the important topics in large document sets and generates a summary which consists of extracts related to identified topics. With respect to the work described in this thesis, the first proposed technique adopts a similar approach to the above by attempting to generate summaries using straightforward classification.

2.5.4 Evaluation measures for Text Summarisation

Steinberger and Ježek [98] presented a taxonomy of summary evaluation measures categorising them as being either: (i) intrinsic and (ii) extrinsic. Intrinsic evaluation is directed at the analysis and comparison of the generated summary with the original document or with a summary generated by a human. Extrinsic evaluation is directed at determining how useful a summary is with respect to a certain domain. Intrinsic evaluation is then subdivided into: (i) text quality and (ii) content evaluation. Text quality evaluation requires that summaries should be: (i) grammatically correct, (ii) non-redundant, (iii) present referential clarity, (iv) have structure and (v) coherence. Content evaluation, on the other hand, is more quantitative and is further subdivided into two categories: (i) co-selection and (ii) content-based. Co-selection considers: (i) precision, (ii) recall, (iii) F-score and (iv) relative utility. Content-based considers: (i) cosine similarity, (ii) unit overlap, (iii) longest common subsequence, (iv) n-gram matching, (v) Pyramids and (vi) Latent Semantic Analysis (LSA) measures. The extrinsic evaluation consists of three measures: (i) document classification, (ii) information retrieval and (iii) question answering.

The labels used with respect to the text summarisation output factors proposed by Afantenos [2] (see Subsection 2.5.1) can also be used to evaluate summaries. Summaries should be: (i) complete, (ii) accurate and (iii) coherent. As suggested in [2], there is still no general consensus among the research community on the criteria that can best be used to evaluate a summary since summarisation has a subjective aspect whereby a

generated summary could be considered to be of good quality by some people but not by others. In some cases domain experts have produced “gold” summaries which may be used as benchmarks. However, three problems can be identified: (i) domain experts may not agree on the characteristics that a “gold” summary must have, (ii) it may be resource expensive to generate such “gold” summaries because many domain experts would be required to agree on the criteria to follow, and (iii) it will be time consuming to generate such summaries because, even if the documents to be summarised are short, a considerable number of them should be manually summarised so that an extensive set of “gold” summaries can be produced.

With respect to the research presented in this thesis, the generated summaries were evaluated by both domain experts and by the author of this thesis in terms of completeness and by considering the taxonomy of summary evaluation measures presented in [98]. The intrinsic evaluation measures taken into account were focused on the quality of the text: (i) grammaticality, (ii) non-redundancy, (iii) referential clarity, (iv) structure, (v) completeness, (vi) accuracy and (vii) coherence. The only extrinsic evaluation measure used was a quantitative one, namely document classification, in other words how well the classification methods performed.

2.6 Summary

In summary, this chapter has presented a literature review of the relevant approaches that are close to the research topics covered by this thesis: (i) Questionnaire Data Mining, (ii) Text Classification and (iii) Text Summarisation. The significance of the research presented in this thesis was explained as well as its location in the landscape of the three aforementioned areas of research. An overview of Questionnaire Data Mining approaches was presented, as well as the most relevant approaches depending on the part of the questionnaire they were aimed at: (i) tabular data, (ii) free text and (iii) tabular data and free text combined. In the context of text classification a review was presented of: (i) feature selection techniques, (ii) relevant approaches and (iii) evaluation measures. Finally, with respect to text summarisation the following was reviewed: (i) the categorisation of the text summarisation techniques, (ii) text summarisation approaches, (iii) text summarisation approaches that use text classification methods and (iv) evaluation measures. The following chapter, Chapter 3 will consider the nature and the preprocessing of the data sets used.

Chapter 3

Evaluation Data Sets and Data Preprocessing

3.1 Introduction

In this chapter the nature of the data sets used to evaluate the work presented later in this thesis is described. Three data sets were considered: (i) SAVSNET (Small Animal Veterinary Surveillance Network), (ii) OHSUMED and (iii) Reuters-21578. Of these data sets the SAVSNET data was the most significant, as this was used as the main motivation for the work described in this thesis, in that it is a questionnaire data set (the other two data sets, as will be seen, are more traditional text data sets). The chapter introduces these data sets and also describes the associated data preprocessing; an important step in the context of knowledge extraction in general, and text summarisation in particular. As explained in Chapter 1, the focus of this thesis is to generate summaries through a process of classification. The quality of the generated summaries thus depends on the quality of the generated classifiers, which in turn depends on the nature of the training sets and of course the nature of the class labels assigned to these documents. Free text documents can be related to single or multiple class labels; in the later case the labels can be structured in a hierarchical manner.

Obtaining questionnaire data sets for research purposes is not a straightforward task for reasons of data confidentiality of the data gathered. The data sets used to evaluate the text classification methods used to generate text summaries, presented later in this thesis were as follows:

- SAVSNET
 - SAVSNET-840-4-FT
 - SAVSNET-840-4-TD+FT
 - SAVSNET-971-3-FT
 - SAVSNET-971-3-TD+FT

- SAVSNET-917-4H
- OHSUMED
 - OHSUMED-CA-3187-3H (Cardiovascular Abnormalities)
 - OHSUMED-AD-3393-3H (Animal Diseases)
- Reuters-21578
 - Reuters-21578-LOC-2327-2H (Locations)
 - Reuters-21578-COM-2327-2H (Commodities)

The naming convention used with respect to the above is as follows. Each name comprises three or four elements: (i) title; (ii) sub-title (may be omitted); (iii) total number of records; and (iv) number of classes, or the number of levels (H) in the case of hierarchical data sets. A hierarchical data set is one that features a hierarchy of class labels. In the context of hierarchical data sets the indicator L will be used to indicate the data and class label pairings at a particular level. In the particular cases of the SAVSNET-840-4 and the SAVSNET-971-3 data sets, which contain both tabular data and free text, the indicator FT will be used where the data set contains only free text, and the indicator TD+FT will be used if it contains tabular data and free text. The OHSUMED and the Reuters-21578 data sets are comprised only by free text. Thus we have the SAVSNET data set SAVSNET-840-4-FT which contains 840 records, 4 class labels and only free text. Alternatively, we have the OHSUMED data set OHSUMED-CA-3187-3H which contains 3,187 records arranged into a hierarchy of three levels. Regarding the Reuters-21578 data sets the “21578” indicates the total number of documents in the original data set and is included in the name along with the actual number of documents in the variation of the data set. Note that in the case of the OHSUMED data sets, the number of records in each level is not the same, the first level consists of all the documents and lower levels subsets of the documents in the first level. In some cases, for the purpose of summary generation, the class labels associated with a particular data set at a particular level in a hierarchy are treated independently. In this case the last element of the data set name indicates the level. For example, SAVSNET-917-1L, OHSUMED-CA-2570-2L or Reuters-21578-LOC-2327-2L.

Thus from the above five hierarchical data sets are used in this research: (i) SAVSNET-917-4H, (ii) OHSUMED-CA-3187-3H, (iii) OHSUMED-AD-3393-3H, (iv) Reuters-21578-LOC-2H and (v) Reuters-21578-COM-2H. The SAVSNET-917-4H features a four level hierarchy, the OHSUMED data sets feature three level hierarchies and the Reuters-21578 data sets feature two level hierarchies. It should also be noted that the OHSUMED and the Reuters-21578 data sets are not questionnaire data sets. However, these data sets were used with respect to the work described in this thesis to

evaluate the application of the proposed text summarisation classification techniques to the generation of text summaries from different types of textual data than just questionnaire data. The hierarchical data sets are only applicable with respect to the hierarchical method presented in Chapter 7, for the other three methods considered in this thesis each level of the hierarchy in the data sets is considered as an independent data set, hence the naming of the data sets indicating the levels to which they belong in their respective hierarchy. The data sets will be described in further detail later in this chapter.

The remainder of this chapter is organised as follows. Section 3.2 presents a general discussion of the preprocessing and representation of the textual and tabular elements of questionnaires, as well as free text in general. The individual data sets (SAVSNET, OHSUMED and Reuters-21578) used in the experiments that were carried out to evaluate the proposed approaches presented in this research, as well as their preprocessing, are then described in detail in Sections 3.3, 3.4 and 3.5 respectively. A summary of the chapter is presented in Section 3.6.

3.2 Data Preprocessing

In this section a general discussion concerning data preprocessing in the context of data mining is presented. Specific details of the preprocessing of the evaluation data sets used with respect to the work described in this thesis is presented in later sections. Although the raw questionnaire data from which it is desired to extract knowledge is expected to be in an electronic format (transcribed and/or collected electronically), it still needs to be preprocessed in order that data can be translated into an appropriate format that will permit the application of classification techniques. Raw data typically contains noisy elements such as corrupted, redundant or incomplete data (data that features missing values), and data not relevant for the knowledge extraction process. Data preprocessing is the first step of the KDD process [41]. Data preprocessing can be time-consuming, but is of great importance in order to improve the quality of the data input to the data mining step. In the context of questionnaire data, and assuming that both the tabular data and the free text elements are of interest, both need to be preprocessed and represented separately. In the following two subsections a general description of how preprocessing is typically carried, with respect to both tabular and free text data, is presented.

3.2.1 Tabular Data

Although this thesis is focused on summarising the free text element of questionnaires, the preprocessing of the tabular element is also considered (although to a lesser extent) in this chapter because, in addition to often being present in questionnaires, it can

provide additional valuable information with respect to the text summarisation process that could enrich the free text to be summarised: (i) by including relevant tabular attributes as words in the free text during the classification process and (ii) by using the output of mining tabular data in conjunction with the output of the free text classification. As in the case of tabular data in general, once collected, tabular questionnaire data is typically stored electronically; in some cases bespoke storage systems are used which may have some built in data analysis functionality. For data mining purposes it is usually necessary to transform the data into a format compatible with the data mining software to be adopted. Common formats include: CSV, SSV, XML, TAB and ARFF, amongst others. Han *et al.* [41] consider that tabular data preprocessing includes techniques such as:

- **Data cleaning:** Removal of noise and fixing of inconsistencies.
- **Data integration:** Integration of data from different sources.
- **Data transformation:** Transformation of data into more suitable formats for the data mining process.
- **Data reduction (Feature selection):** Reduction of the size of the data set, taking into account the most important information, so as to decrease the overall anticipated computational cost.

3.2.2 Free Text

There are many ways to represent free text documents for data mining purposes, two representations that frequently appear in the literature are: (i) the *Vector Space Model (VSM)* and (ii) *Latent Semantic Indexing (LSI)* [23]. In the VSM, free text documents are represented as vectors (one per document); in this context the terms contained in the vectors can be words or phrases. Where words are used this is referred to as the *bag-of-words* representation; where phrases are used this is referred to as the *bag-of-phrases* representation. More specifically, the bag-of-words representation refers to an unordered representation of a text document based on the single words that appear in it. In the bag-of-phrases representation, on the other hand, a text document is represented by the phrases that appear in it. LSI creates an index of the words contained in a text document and makes use of *Singular Value Decomposition (SVD)* to reduce the dimensionality of the data by generating a mapping of the relationships between words and documents. Words and documents that are closely related to each other are put near one another. SVD has been used extensively in the area of statistics and in this context is used to identify how much the words and the documents are related to each other, focusing on the ones that have a close association. Words in a

query are then used to retrieve the related documents based on how near they are to the location represented in the mapping.

The free text contained in questionnaire data can be viewed as a special form of free text in that it features certain characteristics that are typically not found in more standard forms of free text. More specifically, questionnaire free text: (i) tends to be unstructured, (ii) frequently includes misspelled words, (iii) features poor grammar, and (iv) often includes abbreviations and acronyms specific to the domain. As such, questionnaire free text has similarities with e-mail correspondence, mobile phone texts and tweets, as opposed to the free text found in (say) newspaper articles or document collections. The advantage of analysing the free text element of questionnaires, as already noted, is that this free text tends to be much more informative with regard to respondent opinions than the tabular element of questionnaire data.

The bag-of-words representation was adopted, for use with respect to the research described in this thesis, to represent the textual data. The reasons for using this representation were directly related to the nature of questionnaire free text (lack of structure of the free text, misspelled words, poor grammar, abbreviations and acronyms) where by it was not viable to use a representation based on syntax or semantics such as the bag-of-phrases representation or LSI. The main disadvantage of the bag-of-words representation is that the relationship (ordering) between the words is lost. However, this was not considered significant given the nature of the questionnaire free text under consideration.

Text preprocessing tends to adopt some of the tabular data preprocessing techniques together with some additional techniques directed specifically at textual data. Miner *et al.* [74] presents a list of basic text preprocessing steps:

1. **Scope of documents selection:** Deciding whether to use an entire free text document collection, or just some part of it that is considered to be most relevant to a specific text mining application.
2. **Tokenization:** Breaking the text down into single words or tokens, typically using white spaces and punctuation symbols as delimiters.
3. **Stop word removal:** Stop words are common words that are considered to be not significant given a desired application (for example articles such as “the” or “a” are often considered to be stop words) and are thus frequently removed.
4. **Stemming:** Stemming refers to processing text so that, where appropriate, words are reduced to their stem base or root by removing prefixes and suffixes; for example the words “operated”, “operating” and “operates” share the same stem base “operat”.

5. **Spelling normalisation:** The process of automatically correcting misspelled words by using dictionary-based approaches, fuzzy matching algorithms or word clustering.
6. **Sentence boundary detection:** Breaking down the text into sentences.
7. **Case normalisation:** Homogenising the given text, which is typically written in mixed case, into lower or upper case.

The specific preprocessing associated with each of the data sets used for evaluation purposes in this thesis are discussed in the following three subsections.

3.3 Small Animal Veterinary Surveillance Network (SAVS-NET)

In this section the SAVSNET data is described. The section is divided into two subsections. In the first subsection an overview of the SAVSNET data is presented and in the following subsection the associated preprocessing is described.

3.3.1 Description of the SAVSNET data set

As already noted, the questionnaire data used to evaluate the proposed methods described in this thesis was generated as part of the SAVSNET (Small Animal Veterinary Surveillance Network) project [83]. SAVSNET is an initiative that is currently in progress within the Small Animal Teaching Hospital at the University of Liverpool in the UK. The objective of SAVSNET is to provide information on the frequency of occurrence of small animal diseases (mainly in dogs and cats). The project is partly supported by Vet Solutions, a software company whose software is used by some 20% of the veterinary practices located across the UK. Some 30 veterinary practices, all of whom use Vet Solutions’ software, have “signed up” to the SAVSNET initiative.

The SAVSNET veterinary questionnaires comprise a tabular (tick box) section and a free text section. Each questionnaire describes a consultation and is completed by the vet conducting the consultation. The tabular section of the questionnaires includes questions that are associated with general details concerning the consultation (e.g. date, consultation ID, practice ID), while others are concerned with the “patient” (e.g. species, breed, sex) and its owner (e.g. postcode). The free text section of the questionnaires usually comprises notes made by the vet, which typically describe the symptoms presented, the possible diagnosis and the treatment to be prescribed. It is the free text section that we are interested in summarising, although in some cases the free text element of the questionnaires is left blank.

Not all the consultations that take place in a given veterinary practice are applicable with respect to the SAVSNET questionnaire, only some of them are eligible

according to a selection criteria established by SAVSNET. The precise format of the questionnaires was also varied from time to time so as to direct the questionnaire focus on different aspects of veterinary practice. Typically one of the closed questions was changed periodically to reflect a “condition” of interest. These were thus the class attributes that were selected to support the summarisation classifier generation process with respect to this thesis. For example, at one stage, a closed-ended question relating to “vomiting” and “diarrhoea” was included (common conditions found in small animals). The potential answers thus defined a set of class labels. The SAVSNET questionnaires were collected on a continuous basis, but for the experiments carried out in this research only five subsets of the SAVSNET questionnaire corpus were used: (i) SAVSNET-840-4-FT, (ii) SAVSNET-840-4-TD+FT, (iii) SAVSNET-971-3-FT, (iv) SAVSNET-971-3-TD+FT and (v) SAVSNET-917-4H. They were selected with the assistance of domain experts who took into account their interestingness [94] (in terms of Veterinary Science) and the amount of data available. Each of the five selected subsets is considered in some further detail in the following subsections.

3.3.1.1 SAVSNET-840-4-FT

The SAVSNET-840-4-FT data set comprised 840 free text records and 4 class values: (i) Aggression, (ii) Diarrhoea, (iii) Pruritus and (iv) Vomiting. The number of records per class is presented in Table 3.1.

Table 3.1: Number of records per class in SAVSNET-840-4-FT.

Class	Num.	%
<i>Aggression</i>	34	4.05
<i>Diarrhoea</i>	315	37.50
<i>Pruritus</i>	352	41.90
<i>Vomiting</i>	139	16.55
Total	840	100.00

3.3.1.2 SAVSNET-840-4-TD+FT

The SAVSNET-840-4-TD+FT (recall that “TD+FT” stands for “Tabular Data plus Free Text”) data set is an extended version of the SAVSNET-840-4-FT data set which includes selected tabular attributes, in the form of words, in the free text as a way of enriching the generation of text summarisation classifiers. Therefore it has the same number of records and classes as the SAVSNET-840-4-FT data set. Later in this thesis the results generated using both SAVSNET-840-4-FT and SAVSNET-840-4-TD+FT are compared in order to provide an insight into the effect of including tabular data in the classification process. Note that in the SAVSNET case there is a relatively small number of tabular attributes, as opposed to some other questionnaire data sets which

may include a larger number of tabular attributes that may be included in the free text.

3.3.1.3 SAVSNET-971-3-FT

The SAVSNET-971-3-FT data set contained 971 free text records and 3 class values: (i) Diarrhoea, (ii) Vomiting and (iii) Vomiting and diarrhoea. The number of records per class is presented in Table 3.2. Note that the “*Vom & Dia*” (Vomiting and Diarrhoea) class is closely related to the other two classes (“*Diarrhoea*” and “*Vomiting*”), which could lead to a conflict and possible misclassifications. However the domain experts determined that the documents to be labelled with the “*Vom & Dia*” class were the ones where both conditions were presented and they were not taken into account for the other labels to avoid redundancy.

Table 3.2: Number of records per class in SAVSNET-971-3-FT.

Class	Num.	%
<i>Diarrhoea</i>	586	60.35
<i>Vomiting</i>	268	27.60
<i>Vom & Dia</i>	117	12.05
Total	971	100.00

3.3.1.4 SAVSNET-971-3-TD+FT

In a similar manner as with the SAVSNET-840-4-FT data set, the SAVSNET-971-3-TD+FT data set is an extended version of the SAVSNET-971-3-FT data set which also includes selected tabular attributes in the free text. Thus SAVSNET-971-3-TD+FT has the same number of records and classes as the SAVSNET-971-3-FT data set. Again, the data set was generated so as to compare the effect on the classification process of including tabular elements.

3.3.1.5 SAVSNET-917-4H

The SAVSNET-917-4H data set consists of 917 records and several class attributes arranged over 4 different levels in a class hierarchy defined by specific questions variously included in the SAVSNET questionnaires. This data set therefore comprised four sub-data sets associated with each level in the hierarchy: (i) SAVSNET-917-1L, (ii) SAVSNET-917-2L, (iii) SAVSNET-917-3L and (iv) SAVSNET-917-4L. For the non-hierarchical methods presented in this thesis each of these data sets was considered as an independent data set. The distribution of the documents per class, for the four levels in the hierarchy, is shown in Tables 3.3, 3.4, 3.5 and 3.6. Note that the same documents appear in each level of the hierarchy, but with different class labels.

Table 3.3: Number of records per class in SAVSNET-917-1L.

Class	Num.	%
<i>Diarrhoea</i>	536	58.45
<i>Vomiting</i>	248	27.05
<i>Vom & Dia</i>	133	14.50
Total	917	100.00

Table 3.4: Number of records per class in SAVSNET-917-2L.

Class	Num.	%
<i>Haemorrhagic</i>	177	19.30
<i>Not Haemorrhagic</i>	604	65.87
<i>Unknown Severity</i>	136	14.83
Total	917	100.00

Table 3.5: Number of records per class in SAVSNET-917-3L.

Class	Num.	%
<i>First Time</i>	573	62.49
<i>Nth Time</i>	290	31.62
<i>Unknown Occurrence</i>	54	5.89
Total	917	100.00

Table 3.6: Number of records per class in SAVSNET-917-4L.

Class	Num.	%
<i>Less Than One Day</i>	273	29.77
<i>Between Two And Four Days</i>	411	44.82
<i>Between Five And Seven Days</i>	82	8.94
<i>More Than Eight Days</i>	139	15.16
<i>Unknown Duration</i>	12	1.31
Total	917	100.00

3.3.2 Preprocessing of the SAVSNET data set

This subsection describes the preprocessing that was carried out on the SAVSNET free text and tabular questionnaire data and the adopted representations; as well as the merging of relevant table data attributes (in the form of words) with the free text as a means of enriching the free text. With respect to the research described in this thesis both the free text and the tabular part of the questionnaires were represented in a CSV format to ensure compatibility with the open source data mining software used; namely: (i) the LUCS-KDD (Liverpool University Computer Science - Knowledge Discovery in Data) DN (Discretisation/Normalisation) software [16], (ii) the LUCS-KDD TFPC (Total From Partial Classification) Classification Association Rule Mining (CARM) algorithm [18], (iii) the WEKA (Waikato Environment for Knowledge Analysis) machine learning workbench [39] and (iv) the Orange data mining software [21].

As already noted, the SAVSNET data set consisted of questionnaires directed at a particular target population in order to gather information regarding a specific domain. The completion of questionnaires was conducted in either hard or soft form. Either way, answers to open and closed ended questions were typically stored in spreadsheets, databases or CSV files. An example CSV file fragment extracted from the SAVSNET-840-4 questionnaire return data is presented in Figure 3.1 and summarised in tabular form in Table 3.7. The “class labels” were defined by the domain experts who devised the questionnaires and were encoded as a separated attribute (attributes in the case of multiple class labels). The first line in Figure 3.1 is the header to the CSV file and indicates the names of the attributes that may be present in each record. Missing attribute values are indicated by a space (consecutive commas with no values recorded between them). In the examples presented in Figure 3.1 the last attribute (ConsultationNotes) is the free text attribute of interest with respect to the work described in this thesis. The vertical bar symbol (|) used in Figure 3.1 is a delimiter between fragments of text.

The data presented in Figure 3.1 is summarised in tabular form in Table 3.7. For ease of understanding, only six attributes are shown in Table 3.7: ID, Species, Breed, Sex, Colour and ConsultationNotes. The first five attributes are answers to closed-ended questions where values can be numeric, nominal or boolean (yes/no). The last attribute (ConsultationNotes), and that of principal interest with respect to this thesis, contains the answer to an open-ended question concerned with notes made by a vet during a consultation. In Table 3.7 the free text presented in the “ConsultationNotes” column is truncated due to the length of the entire free text, however the full text is available from Figure 3.1. The table does not include the class label associated with each record because at this stage the class labels had not been explicitly determined (the class labels were determined later from specific tabular attributes). It should also be noted that the information presented in Figure 3.1 and Table 3.7 will be referred back to later in this chapter.

ID,IP,What,Date,Time,Survey,Species,Breed,DOB,Sex,Neutered,PostCode,PracticeID,AnimalID,Weight,Colour,MicroChipped,Insured,Deceased,TransactionID,SurgeryID,Data,Occurrence,Haemorrhagic?,Duration,Treatment,Diagnostic,Haem Vom?,Haem Dia?,Affected Area,Has Presented Before, Diarrhoea,Symptoms,Vomiting,ConsultationNotes

34039,83.148.170.241,Examination by Vet,20100526,14:10:07,Vomit & Diarrhoea,Feline,DSH,19970526,UnKnown,Yes,,2014,19034,4.4,Tortois + Whi,No,No,No,9,1,,1st presentation of vomiting,Vomit not haemorrhagic,Unknown,Treatment not listed,,,,,,,,,"Registration form completed | V+ on sunday, O reports often V+ with furballs. O also reports constipation, with faeces, U+ dark. Appetite depressed, drinking small amounts. No blood in V+, No D+. | Weight 4.40 kgs. | On exam: Abdo extremely distended. BCS 1, Pale mm prolonged skin tent. Offered investigation & intensive tx/pts. | SAVSNET Vomit and Diarrhoea | survey completed | Examination by Vet | Est 59.61 /Cat | Form (text/euth.def) | euthanasia -cat | Needs pts, O does not have funds. MH adv AW/RSPCA. Signed consent form. | "

34057,80.177.110.171,Consultation Dog,20100518,17:03:30,,Canine,Podenco Ibicenco,20020819,Male,Yes,LE3 8BH,2006,62439,33.5,Brown,Yes,No,No,30000214,3,,1st presentation of diarrhoea,Haemorrhagic diarrhoea,Dia <1 day,"Antibiotic therapy,NSAIDS,worming",No diagnostic tests performed,,,,,,,,,"SAVSNET Vomit and Diarrhoea | survey completed | Consultation Dog | Inject 2.50 mls RIMADYL INJ 20ML | Inject 3.30 mls BETAMOX LA | mucus D+ and fresh blood last 24hrs, ok in self. adv worm and light diet until faeces N, TSA if not much imp within 48hrs or if worsesn/ recurs. | "

34077,80.177.110.171,Consultation Dog,20100518,17:29:31,,Canine,Labrador,19980915,UnKnown,Yes,LE6 0DY,2006,19796,30,Black,Yes,Yes,No,30000216,3,,1st presentation of diarrhoea, Haemorrhagic diarrhoea,Dia 2-4 days,"Antibiotic therapy,Diet modified,NSAIDS, wormer",No diagnostic tests performed,,,,,,,,,"SAVSNET Vomit and Diarrhoea | survey completed | Consultation Dog | Inject 3.10 mls BETAMOX LA | Inject 2.50 mls RIMADYL INJ 20ML | fesh blood and mucus d+ last 3 days, well in self still eating, hydrtated, clin NAD except sl tense abdo, palp NAD, TN. light diet and wormer, TSa if necc. | 6pack DRONTAL PLUS Give one tablet per 10kgBW in food | 2 packs ATOPICA 100MG as directed | pack ATOPICA 50MG DOGS7.5-36KG | "

34295,82.153.142.237,CONSULTATION/EXAMINATION,20100518,17:46:05,,canine,Eng Springer Spaniel,20090518,Male,No, TQ1 1NB,2011,39848,21.9,Liver/white,Yes,No,No,30000024,3,,1st presentation of diarrhoea,Diarrhoea not haemorrhagic,Dia 5-7 days,"Diet modified,kaogel, electrolyte solution,24hr dexamethasone",No diagnostic tests performed,,,,,,,,,"SAVSNET Vomit and Diarrhoea | survey completed | CONSULTATION/EXAMINATION | colitic type d+ past 4d, less frequent, o straved then bland food, ?weight loss, appetite wnl, drinking wnl. inc vol, liquid, 1x day few specks bld, straining. no v+. eating grass. otherwise ok. CE mm,crt wnl, BAR, heart/lungs | wnl, abd nad. Adv cont bland food, tx symptomatically initially but faecal sample ini | Injection 1.10 x colvasone inj 50ml Batch:9253-91 Exp: 06.11 | Supply Glutalyte add to 500ml water Batch:9133-43 Exp: 03.11 | Supply 100ml KAOGEL VP [Discount of 10.0%] Batch:5131738 Exp: 12.13 | weight 21.60 kgs. | "

34309,62.49.78.145,Consultation Dog,20100510,18:33:47,,Canine,Staffordshire Bull Terrier,20060906,Male,No,LE3 2XN,2006,78560,23.95,Brindle & Whi,Yes,No,No,40000009,4,,1st presentation of diarrhoea,Haemorrhagic diarrhoea,"Dia <1 day,Unknown","Antibiotic therapy, Antibiotic therapy,Diet modified",enema,,,,,,,,,"yesterday O gave him a bone this am bloody watery D++ also V++ this am was a proper dog bone also a little depressed can feel shards of bone in rectum removed a large piece lots of smaller pieces in rectum adv tx | abs rim & cat micalax TSA tomorrow at g/f am if better just cont abs ini admit ga enema | Weight = 23.95 kgs. | SAVSNET Vomit and Diarrhoea | survey completed | Consultation Dog | Inject 1.90 mls RIMADYL INJ 20ML | Inject 1.30 mls SYNULOX RTU | MICRALAX | "

Figure 3.1: Example fragment of raw questionnaire data in CSV format.

Table 3.7: Example fragment of raw questionnaire data in tabular form.

ID	Species	Breed	Sex	Colour	ConsultationNotes
34039	Feline	DSH	Unknown	Tortois + Whi	Registration form completed V+ on sunday, O reports often V+ with furballs. O also reports constipation...
34057	Canine	Podenco Ibicenco	Male	Brown	SAVSNET Vomit and Diarrhoea survey completed Consultation Dog Inject 2.50 mls RIMADYL INJ 20...
34077	Canine	Labrador	Unknown	Black	SAVSNET Vomit and Diarrhoea survey completed Consultation Dog Inject 3.10 mls BETAMOX LA ...
34295	canine	Eng Springer Spaniel	Male	Liver/white	SAVSNET Vomit and Diarrhoea survey completed CONSULTATION/ EXAMINATION colitic type d+ past...
34309	Canine	Straffordshire Bull Terrier	Male	Brindle & Whi	yesterday O gave him a bone this am bloody watery D++ also V++ this am was a proper dog bone also a little...

Although it is not the case in the free text presented in Figure 3.1, there were SAVSNET records that did not contain any free text, in which case the free text element was represented as an empty string (“”). Each “chunk” of free text was referenced using the associated ID number. It should be noted that the free text examples given in Figure 3.1 include abbreviations and acronyms related to the domain of Veterinary Science; for example, abbreviations such as “V+” or “D+” stand for “vomiting” and “diarrhoea” respectively, while the abbreviation “O” stands for “owner”. Some words and phrases, like “Consultation Dog”, are of an administrative nature and for many purposes may not be of any relevance with respect to the extraction of useful information. In addition, as indicated by the sample text given in the figure, it is often the case that free text questionnaire responses include typing, spelling and grammar errors that serve to hinder the preprocessing. Depending on the type of analysis to be carried out on the free text, numbers and symbols can be considered as either: (i) useful features that could add information or (ii) simply as noise. Note that for a non-domain expert it is difficult to have a clear and complete understanding of the free text without recourse to domain experts.

As has been mentioned previously in this thesis, processing and analysing free text is not as straightforward as in the case of tabular data, mainly due to the complex way in which ideas and information are represented in free text. For example, in the context of the SAVSNET data sets, the species of a small animal that was examined by vets in a clinic can be given by either a nominal value, such as “feline”, under the “Species” attribute in the tabular part of a questionnaire; or in the free text as: “Owner brought a small cat to Leahurst clinic for examination”, where “cat” implies the species of the animal is “feline”. It is clear that a significant amount of information, apart from simply an animal’s species, can be extracted from the free text element of

questionnaire data; for example, the purpose of the consultation or the clinic to which the cat was brought to.

Overall, and based on the general free text preprocessing steps suggested by Miner *et al.* [74] presented above, the SAVSNET questionnaire free text used within the context of this thesis was preprocessed as follows:

1. **Scope of documents selection:** The entire collection of free text “documents” was selected (as opposed to some sample).
2. **Case normalisation:** Words were converted to lower case to standardise them. Since the techniques used in this research are not NLP-based, this conversion did not affect the data analysis.
3. **Numbers, symbols and stop words removal:** Numbers and symbols were removed from the free text because they were considered to represent noise; in a bag-of-words representation only dictionary words (even if they are misspelled) are desired. Note that in some domains there are particular numbers or symbols that when put together with a word or a letter could have significance for the data analysis. For example, in the context of the SAVSNET data, “v++” indicates the presence of significant vomiting in animals. The assistance of a domain expert was required so as to identify such symbols. Stop words were also removed.
4. **Tokenization:** The text was tokenized into single words using white spaces and punctuation symbols as delimiters.
5. **Stemming:** To overcome the twin issues of misspelled words and poor grammar in the free text, an implementation of the Porter Stemming algorithm [106] was adopted. In the early stages of this research misspelled words were addressed by using a spell checker, however this proved to be unuseful, as many words that shared the same stems were assigned to different and unrelated “corrected” words. Although applying stemming transforms the words and reduces the meaning contained in them, since the text summarisation techniques presented in this thesis were based on text classification and not on the extraction of sentences or other content of the free text (as in the case of more standard text summarisation techniques), stemming was found to not affect the overall process of summary generation.
6. **Feature selection:** In a similar manner to the preprocessing of tabular questionnaire data (see above), feature selection techniques were applied to select the most relevant features (words) contained in the free text. Note that free text usually contains considerably more features than tabular data. The feature selection was directed at two main objectives: (i) to reduce the dimensionality of

the data and hence save on the computational resource required to process the data, and (ii) to select the most relevant features so as to improve the effectiveness of the desired classification/summarisation. In the case of the SAVSNET data, domain experts were used to advise on the validity of selected features. For comparison purposes two alternative feature selection techniques were considered: (i) Term Frequency-Inverse Document Frequency (TF-IDF) [63] plus Chi-squared [113] and (ii) TF-IDF plus Correlation-based Feature Selection (CFS) [40].

For the semi-automated method presented in Chapter 6 the preprocessing of the free text consisted only of the removal of numbers and symbols while keeping phrase delimiters such as commas, semicolons and full stops in order to have a clean but coherent free text from which the user could identify relevant phrases. The resulting preprocessed text was thus saved in a different file format to that used for the other methods.

As noted above, the merging of relevant tabular data attributes, in the form of words, with the free text is also investigated in this thesis using extended versions of the SAVSNET-840-4 and SAVSNET-971-3 data sets (SAVSNET-840-4-TD+FT and SAVSNET-971-3-TD+FT respectively) as it was conjectured that this could be a means of enriching the free text. The desired merging required that the tabular attributes that were of interest contained nominal values. Boolean values represented as nominal values could also be taken into account. For example, the SAVSNET-840-4 data set includes an attribute “neutered” which can take the values “no” or “yes” which can be interpreted as “unneutered” and “neutered” respectively. This merging of tabular data with the free text was viable because the proposed techniques used to extract summaries from the free text did not rely on syntax or semantics. Figure 3.2 shows a fragment of the SAVSNET-840-4-TD+FT data set after the free text had been partially preprocessed (text was put into lower case; numbers, symbols and stop words were removed; and the text associated with each record translated into a bag-of-words vector) and merged with four tabular attributes added before the bag-of-words words (first four “words” for the free text attribute). The eight tabular attributes were: “species”, “breed”, “sex”, “neutered”, “colour”, “microchipped”, “insured” and “deceased”. It was conjectured, with the exception of the attribute “sex” (third word in the text example), that these attributes might add more relevant information that could improve the text classification and consequently the summaries generated. In the case of the “sex” attribute the contribution was not as clear because there were many unknown values. With respect to the SAVSNET data there were few tabular attributes that could be appropriately merged with the free text. The data presented in Figure 3.2 also includes a class attribute (last attribute for each record highlighted in bold font). The class labels were obtained using specific tabular attributes that indicated the presence of a certain condition of interest.

id	freeText	class
34039	feline dsh sexUnknown neutered tortois whi notMicrochipped uninsured notDeceased v+ sunday reports v+ furballs reports constipation feaces dark appetite depressed drinking small amounts blood exam abdo expremely distended bcs pale mm prolonged skin tent offered investigation intensive tx pts vet est cat form text euth def euthanasia cat pts funds mh adv aw rspca signed consent form,	VomitingClass
34057	canine podencoibicenco male neutered brown microchipped uninsured notDeceased ect rimadyl ect betamox la mucus d+ fresh blood adv worm light diet faeces tsa imp worsesn recurs,	DiarrhoeaClass
34077	canine labrador sexUnknown neutered black microchipped insured notDeceased ect betamox la ect rimadyl fesh blood mucus d+ well eating hydrtated clin nad sl tense abdo palp nad tn light diet wormer tsa necc pack drontal bw food packs atopica directed pack atopica dogs,	DiarrhoeaClass
34295	canine engspringerspaniel male unneutered liver white microchipped uninsured notDeceased colitic type d+ frequent straved bland food loss appetite wnl drinking wnl liquid day specks bld straining eating grass ce mm crt wnl bar heart lungs wnl abd nad adv cont bland food tx symptomatically initially faecal sample ini colvasone batch exp supply glutalyte add water batch exp supply kaogel vp discount batch exp,	DiarrhoeaClass
34309	canine staffordshirebullterrier male unneutered brindle whi microchipped uninsured notDeceased yesterday bone bloody watery d++ v++ proper dog bone depressed feel shards bone rectum removed large piece lots smaller pieces rectum adv tx abs rim cat micalax tsa tomorrow better cont abs ini admit ga enema ect rimadyl ect synulox rtu micralax,	DiarrhoeaClass

Figure 3.2: Example fragment of preprocessed free text merged with tabular attributes.

Inspection of the five records shown in Table 3.7 indicates that they include: (i) case inconsistencies (“Canine” and “canine”), (ii) unknowns (Sex column) and (iii) inclusion of symbols in text (“Tortois + Whi”, “Liver/white”, “Brindle & Whi”). It is a common feature of tabular questionnaire data that missing values and redundancies are included. The general steps used to preprocess the SAVSNET tabular data in this research were as follows:

1. **Case normalisation:** String data types were converted to lower case to obtain a consistent data representation.
2. **Number and symbol removal:** Numbers and symbols contained in nominal attributes were removed. Symbols and white spaces that indicated the separation between two words within a string of characters (for example “White Tortois”) were replaced with a unique symbol (“/”).
3. **Missing values handling:** There are many strategies to address the issue of missing values [41], examples include: (i) ignoring the record, (ii) filling the missing value manually, (iii) using a global constant to indicate the missing value, (iv) using the attribute mean to assign a value, (v) using the attribute mean for all the samples belonging to the same “class” of the selected record and (vi) using the most probable value to fill in the missing value. The fourth strategy, using the attribute mean to assign a value, was used in this research.
4. **Feature selection:** Not all the tabular attributes were of interest to the data mining process, thus feature selection techniques (Chi-squared or CFS) were applied to identify which attributes were the most significant with respect to the associated classes (which were the best discriminators); irrelevant attributes could then be discounted. Domain experts were also used to validate the selected attributes (in some cases they suggested additional attributes to be included).
5. **Discretisation:** Finally, in the case of numeric continuous attributes it was necessary to discretise them; generating ranges of values from continuous data, so that continuous values could be treated as nominal attributes.

Figure 3.3 shows a fragment of the tabular element of the SAVSNET-840-4 data set. Note that the most relevant attributes, as confirmed by the domain experts, are shown. The schema for the data is presented in the first line.

3.4 OHSUMED

In this section the OHSUMED data set is described. As in the case of the foregoing section, this section comprises two subsections. In the first subsection the OHSUMED

id	species	breed	sex	neutered	colour	microchipped	insured	deceased	class
34039	feline	dsh	unknown	yes	tortois&whi	no	no	no	VomitingClass
34057	canine	podencoibicenco	male	yes	brown	yes	no	no	DiarrhoeaClass
34077	canine	labrador	unknown	yes	black	yes	yes	no	DiarrhoeaClass
34295	canine	engspringerspaniel	male	no	liver&white	yes	no	no	DiarrhoeaClass
34309	canine	staffordshirebullterrier	male	no	brindle&whi	yes	no	no	DiarrhoeaClass

Figure 3.3: Example fragment of tabular questionnaire data in CSV format.

data sets are introduced, in the following subsection the associated preprocessing is described.

3.4.1 Description of the OHSUMED data set

The OHSUMED data set was generated by Hersh *et al.* [44] and comprises 348,566 references (titles and/or abstracts) from the MEDLINE database, which in turn contains around 19 million citations for biomedical literature, including journals and books¹. The references covered by OHSUMED are from 270 medical journals collated over a five-year period (1987-1991). The field definitions of the OHSUMED data set are: (i) a sequential identifier, (ii) a MEDLINE identifier (UI), (iii) human-assigned MeSH (Medical Subject Headings) terms (MH), (iv) title (TI), (v) publication type (PT), (vi) abstract (AB), (vii) author(s) (AU) and (viii) source (SO). The field definitions that were used to extract the data sets used with respect to the work described in this thesis were: (i) the MEDLINE identifier (which constitutes a unique ID for each document), (ii) the MeSH terms (which define the class labels of the documents), and (iii) the title and the abstract (which constituted the desired free text).

Two subsets of OHSUMED were used in the experiments described later in this thesis: (i) OHSUMED-CA-3187-3H (CA stands for Cardiovascular Abnormalities) and (ii) OHSUMED-AD-3393-3H (AD stands for Animal Diseases). Each will be described in further detail in the following two subsections. The MeSH terms contained in the documents were used based on the “MeSH Tree Structures” defined by the U.S. National Library of Medicine², specifically the ones concerning diseases. The reasons why the OHSUMED data set was selected were because: (i) its usage is frequently reported in the literature for the purpose of evaluating text classification techniques [10, 61, 110], (ii) the considerable number of documents and classes available in comparison to the SAVSNET data set and (iii) the hierarchical organisation of the classes which is given by the MeSH tree structures and which therefore provides for the evaluation of the

¹http://www.nlm.nih.gov/databases/databases_medline.html

²<http://www.nlm.nih.gov/bsd/disted/mesh/tree.html>

hierarchical summarisation method presented in Chapter 7. For both OHSUMED-CA-3187-3H and OHSUMED-AD-3393-3H, each level of their respective hierarchies was considered as an independent data set for the evaluation of the non-hierarchical methods presented later in this thesis. Unlike the subsets of the SAVSNET-917-4H (described above) and the Reuters-21578 data sets (described below in Section 3.5), the subsets of OHSUMED used in this research do not contain the same number of documents per level in the hierarchy of documents because the documents were already organised in a hierarchical way in which not all documents have a representation at all levels of the hierarchy. As a consequence, the length of the summaries obtained from the OHSUMED data sets varied depending on the number of levels of the hierarchy in which the documents to be summarised were represented.

3.4.1.1 OHSUMED-CA-3187-3H (Cardiovascular Abnormalities)

The reason for choosing the OHSUMED-CA-3187-3H (Cardiovascular Abnormalities) related documents to generate a data set was because it is a subset of the “Cardiovascular Diseases” data set, which has been widely used in the literature [13, 38, 62, 68] and which contains other related topics such as “Heart Diseases” and “Vascular Diseases”. The data set was hierarchically organised into three levels. Table 3.8 shows the number of classes and documents per level. The names of the classes, the MeSH tree codes and the number of documents per class for each level in the hierarchy, are presented in detail in Tables A.7, A.8 and A.9 of Appendix A. The second level of the OHSUMED-CA-3187-3H hierarchy has a larger number of classes with respect to the first and third levels.

Table 3.8: Number of classes per level in the OHSUMED-CA-3187-3H hierarchy.

Level	Data set	Number of classes	Number of documents
First	OHSUMED-CA-3187-1L	2	3,187
Second	OHSUMED-CA-2570-2L	16	2,570
Third	OHSUMED-CA-834-3L	7	834

3.4.1.2 OHSUMED-AD-3393-3H (Animal Diseases)

The OHSUMED-AD-3393-3H (Animal Diseases) data set was chosen due to the relative familiarity with the subject matter gained by the author while working with the SAVSNET data. This data set was also hierarchically organised in three levels. Table 3.9 presents the number of classes and documents per level. Again, the names of the classes, the MeSH tree codes and the number of documents per class for each level in the hierarchy, are presented in detail in Tables A.10, A.11 and A.12 of Appendix A. In this case the first and second levels of the OHSUMED-AD-3393-3H hierarchy have a larger number of classes than the third level.

Table 3.9: Number of classes per level in the OHSUMED-AD-3393-3H hierarchy.

Level	Data set	Number of classes	Number of documents
First	OHSUMED-AD-3393-1L	34	3,393
Second	OHSUMED-AD-569-2L	26	569
Third	OHSUMED-AD-292-3L	9	292

3.4.2 Preprocessing of the OHSUMED data set

This subsection describes how the OHSUMED data set was preprocessed. As already noted, the OHSUMED data set consisted of references (titles and/or abstracts) from the MEDLINE database, therefore the free text in this data set was different from questionnaire free text in that: (i) it was structured, (ii) there was usually no misspelled words and (iii) correct grammar was typically used. On the other hand, such free text shares some of the challenging characteristics of questionnaire free text: (i) use of abbreviations and (ii) acronyms specific to the domain. As in the case of the SAVSNET free text, the OHSUMED free text was represented in a CSV format to ensure compatibility with the open source data mining software used: LUCS-KDD DN [16], LUCS-KDD TFPC [18], WEKA [39] and Orange [21]. The class labels were already defined in OHSUMED according to MeSH terms, so there was no need to generate them using tabular attributes as in the case of the SAVSNET data. Figure 3.4 shows two raw sample records taken from the OHSUMED data set.

Table 3.10 shows the schema for the OHSUMED data set. Each OHSUMED record can be related to one or more MeSH terms, with respect to the work described in this thesis only the first MeSH term is used as the corresponding class label. As previously stated, the field definitions that are used in the extracted data sets were: the MEDLINE identifier (which constitutes a unique ID for each document), the MeSH terms (which define the class labels for the documents), and the title and the abstract (the free text element). According to the MeSH term selected per record, the associated MeSH code is identified within the MeSH tree structure/hierarchy and used to identify the level in the hierarchy with which records are associated, as well as the hierarchical relations between the classes (parent and child nodes).

Note that given the large number of categories in the MeSH hierarchy (there are 16 main MeSH categories), only two subcategories under the “Diseases” main MeSH category (which in turn contains 26 subcategories) were selected and used to generate the two subsets of OHSUMED: (i) OHSUMED-CA-3187-3H (Cardiovascular Abnormalities) and (ii) OHSUMED-AD-3393-3H (Animal Diseases). Again, based on the general free text preprocessing steps suggested by Miner *et al.* [74], the free text of both the OHSUMED data sets was preprocessed as follows:

1. **Scope of documents selection:** As in the case of the SAVSNET data set, the

.I 237240
.U
90296987
.S
Am Heart J 9010; 120(1):103-9
.M
Adolescence; Aortic Coarctation/PP/*SU; Child; Child, Preschool; Diastole/*PH; Echocardiography;
Echocardiography, Doppler; Electrocardiography; Female; Heart Ventricle/PH; Hemodynamics/PH;
Human; Male; Myocardial Contraction/*PH; Systole/*PH.
.T
Altered systolic and diastolic function in children after "successful" repair of coarctation of the aorta.
.P
JOURNAL ARTICLE.
.W
We investigated whether left ventricular (LV) structural or functional abnormalities persist in children
on long-term follow-up after successful correction of coarctation of the aorta. Two-dimensional
directed M-mode and Doppler echocardiographic examinations were performed in 11 such subjects
and 22 age-matched control subjects. Digitized tracings were made from M-mode recordings of the
LV and Doppler mitral valve inflow recordings to measure septal, posterior wall, and LV dimensions,
LV mass, shortening fraction, peak shortening and lengthening velocities, diastolic filling time,
peak E velocity, peak A velocity, and velocity time integrals. Despite group similarities in age,
body size, and systolic blood pressure, greater fractional shortening ($p = 0.0001$), indexed peak
shortening velocity (p less than 0.001), and greater LV mass index (p less than 0.05) were seen
in the coarctation group in the face of lower LV wall stress ($p = 0.0001$). LV mass index correlated
with the resting arm-leg gradient, which ranged from -4 to +10 mm Hg. The coarctation group
had decreased early filling (p less than 0.006) with compensatory increased late diastolic filling (p
less than 0.05). Diastolic filling abnormalities were prominent in the older coarctation subjects and
were related to both systolic blood pressure (p less than 0.001) and LV mass index (p less than
0.01). Despite apparently successful repair of coarctation of the aorta, persistent alterations in both
systolic and diastolic LV function and LV mass are present in children at long-term follow-up, which
are related to the resting arm-leg gradient.(ABSTRACT TRUNCATED AT 250 WORDS)
.A
Moskowitz WB; Schieken RM; Mosteller M; Bossano R.
.I 144583
.U
89267465
.S
Surgery 8909; 105(6):801-3
.M
Arteriovenous Fistula/*ET/RA; Brachiocephalic Veins/*; Case Report; Catheterization/*AE;
Catheters, Indwelling/AE; Female; Human; Middle Age; Subclavian Artery/*.
.T
Subclavian artery-to-innominate vein fistula: a case caused by subclavian venous catheterization.
.P
JOURNAL ARTICLE; REVIEW; REVIEW, TUTORIAL.
.W
Arteriovenous fistulas caused by subclavian vein catheterization occur rarely. Most subclavian vein
catheters are inserted through an infraclavicular subclavian venipuncture with passage of a vessel
dilator and peel-away sheath over a guidewire. We report a previously undescribed complication of
this technique, namely, a right subclavian artery-to-right innominate vein fistula. The mechanism of
injury was perforation through the opposing walls of the respective vein and artery due to stiffness
of the vessel dilator that could not negotiate the curve from the subclavian vein to the innominate
vein. Measures to avoid this complication are described.
.A
Pabst TS 3d; Hunter GC; McIntyre KE; Parent FN 3d; Bernhard VM.

Figure 3.4: Example of raw OHSUMED data.

Table 3.10: Field definitions for OHSUMED data set.

Code	Meaning
I	Sequential identifier
U	MEDLINE identifier
S	Source
M	MeSH terms
T	Title
P	Publication type
W	Abstract
A	Author(s)

entire collection of free text “documents” was selected for both subsets.

2. **Case normalisation:** As before, words were converted to lower case.
3. **Numbers, symbols and stop words removal:** Numbers and symbols were again removed from the free text because they were considered to represent noise. However, unlike the SAVSNET data, there were no particular numbers or symbols that were considered significant when put together with words or letters. Stop words were also removed.
4. **Tokenization:** The text was again tokenized into single words using white spaces and punctuation symbols as delimiters.
5. **Stemming:** Stemming was used to reduce the dimensionality of the data set and to associate words related to the same concepts or ideas. Again, the Porter Stemming algorithm was adopted [106].
6. **Feature selection:** The same two feature selection techniques used for the SAVSNET data were applied (so as to facilitate comparisons): (i) Chi-squared and (ii) CFS. However, in this case domain experts were not used to validate the selected features.

As in the case of the SAVSNET data, a different preprocessing was carried out with respect to the semi-automated method presented in Chapter 6: numbers and symbols were removed but phrase delimiters (commas, semicolons and full stops) were maintained in order to have a clean but coherent free text from which a user could identify relevant phrases.

An example of some partially preprocessed OHSUMED-CA-3187-3H free text (lower case only; numbers, symbols and stop words removed; bag-of-words vector space format) is shown in Figure 3.5. Since in most cases there were several class labels and most of them had long names, a set of codes was used to easily identify the class label and the level in the hierarchy to which it was related (e.g. AA - First level, first class

label, BC - Second level, third class label, and so on). For example, in Figure 3.5 the record with ID “90296987” has “AA” as the class label, which in turn corresponds to the class “Congenital Heart Defects” and to the MeSH Tree Code “C14.240.400”. In a similar manner, record “89267465” has “AB” as the class label corresponding to the class “Vascular Malformations” and to the MeSH Tree Code “C14.240.850”. Note that at the stage of summary generation the original class labels were used.

id,freeText,class
<p>90296987,altered systolic diastolic function children successful repair coarctation aorta investigated left ventricular lv structural functional abnormalities persist children long term follow successful correction coarctation aorta dimensional directed mode doppler echocardiographic performed subjects age matched control subjects digitized tracings mode recordings lv doppler mitral valve inflow recordings measure septal posterior wall lv dimensions lv mass shortening fraction peak shortening lengthening velocities diastolic filling time peak velocity peak velocity velocity time integrals despite group similarities age body size systolic blood pressure greater fractional shortening indexed peak shortening velocity greater lv mass coarctation group face lower lv wall stress lv mass correlated resting arm leg gradient ranged mm hg coarctation group decreased early filling compensatory increased late diastolic filling diastolic filling abnormalities prominent older coarctation subjects systolic blood pressure lv mass despite successful repair coarctation aorta persistent alterations systolic diastolic lv function lv mass children long term follow resting arm leg gradient abstract truncated,AA</p>
<p>89267465,subclavian artery innominate vein fistula case caused subclavian venous catheterization arteriovenous fistulas caused subclavian vein catheterization occur rarely subclavian vein catheters inserted ough infraclavicular subclavian venipuncture passage vessel dilator peel sheath guidewire report undescribed complication technique subclavian artery innominate vein fistula mechanism ury perforation ough opposing walls respective vein artery stiffness vessel dilator negotiate curve subclavian vein innominate vein measures avoid complication described,AB</p>

Figure 3.5: Example of partially preprocessed OHSUMED data.

3.5 Reuters-21578

In this section the third data set, the Reuters data set, is described. As in the case of the previous two sections, this section is divided into two subsections. The first introduces the data set and gives some background, while the second describes the preprocessing that was applied.

3.5.1 Description of the Reuters-21578 data set

The Reuters-21578 data set is comprised 21,578 English language news stories produced by the Reuters news agency and is one of the most used data sets found in the text classification literature [24, 99, 103]. The text documents within the Reuters-21578 data set are related to 135 classes and were originally collected and labelled by Carnegie Group Inc. and Reuters, Ltd. in 1987³. Around 1991 a first version of the Reuters data set, which was called Reuters-22173, was generated by Lewis and Shoemaker at

³<http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt>

the Center for Information and Language Studies at the University of Chicago³. In 1996 Lewis and Finch generated the Reuters-21578 data set³, which was an improved version of the Reuters-22173 data set (duplicated elements were removed). Unlike the SAVSNET-917-4H and the OHSUMED data sets, the Reuters-21578 data set is not explicitly organised in a hierarchical manner; however, the relation between the 135 classes can be used to generate two hierarchies of classes. The Reuters-21578 data set is represented in 22 files with 21 of them containing 1,000 documents and the remaining one 578 documents. The files are in SGML (Standard Generalized Markup Language) format, which is used for defining generalised markup languages for documents. The parts of the documents are indicated by tags, the most relevant being: (i) *< TOPICS >* (what the document is about, indicated by a list of categories), (ii) *< DATE >* (the date and time of the document), (iii) *< PLACES >* (the place(s) related to the document), (iv) *< PEOPLE >* (names of people involved in the document), (v) *< EXCHANGES >* (similar to TOPICS but with categories related to stock exchanges, e.g. nasdaq), (vi) *< COMPANIES >* (names of companies related to the document), (vii) *< TITLE >* (the title of the document), (viii) *< DATELINE >* (the date of the document) and (ix) *< BODY >* (the free text of the document).

In a similar manner to that reported in [24] and [99], but using different criteria, two hierarchies of classes were generated constituting two subsets of Reuters-21578: (i) Reuters-21578-LOC-2327-2H and (ii) Reuters-21578-COM-2327-2H. Each will be described in the following two subsections. As in the case of the SAVSNET-917-4H and the OHSUMED data sets, with respect to the methods presented in this thesis, each level of each hierarchy was considered as an independent data set with respect to the non-hierarchical methods and as a whole data set for the hierarchical method. As in the case of SAVSNET-917-4H, the number of documents per level in each hierarchy of classes is the same. Similarly to OHSUMED, the main two reasons for using Reuters-21578 with respect to the work described in this thesis were its frequent use in the literature to evaluate text classification techniques and the large number of documents and classes it contains in comparison with the SAVSNET data set. The names of the classes, and the number of documents per class for each level in the Reuters-21578 hierarchies of classes, are presented in detail in Appendix A.

3.5.1.1 Reuters-21578-LOC-2327-2H

The Reuters-21578-LOC-2327-2H hierarchy of classes described the geographical locations for individual news stories. The first level of the hierarchy is related to the region(s) or continent(s) with which each news story was related. The second level of the hierarchy is related to the country or countries to which each news story was related. Table 3.11 presents the number of classes and documents per level. The names of the classes and the number of documents per class for each level in the hierarchy are

presented in detail in Tables A.13 and A.15 of Appendix A.

Table 3.11: Number of classes per level in the Reuters-21578-LOC-2327-2H hierarchy.

Level	Data set	Number of classes	Number of documents
First	Reuters-21578-LOC-2327-1L	14	2,327
Second	Reuters-21578-LOC-2327-2L	92	2,327

3.5.1.2 Reuters-21578-COM-2327-2H

The Reuters-21578-COM-2327-2H hierarchy of classes described the topics to which the news stories reported on. The first level of the hierarchy is related to the types of commodities to which the news stories report on. The second level of the hierarchy is related to the individual commodities to which the news stories report on. Table 3.12 presents the number of classes and documents per level. The names of the classes and the number of documents per classes for each level in the hierarchy are presented in detail in Tables A.14 and A.16 of Appendix A.

Table 3.12: Number of classes per level in the Reuters-21578-COM-2327-2H hierarchy.

Level	Data set	Number of classes	Number of documents
First	Reuters-21578-COM-2327-1L	5	2,327
Second	Reuters-21578-COM-2327-2L	52	2,327

3.5.2 Preprocessing of the Reuters-21578 data set

This subsection describes how the Reuters-21578 data set was preprocessed. As already noted, Reuters-21578 is consisted of news articles, this a different category of free text than that associated with SAVSNET (questionnaires) or OHSUMED (abstracts from academic papers). However, Reuters-21578 is similar to OHSUMED, and different from SAVSNET, in that it contains structured free text. Thus the preprocessing is similar to that carried out for the OHSUMED data. As in the case of the SAVSNET and OHSUMED data, the Reuters-21578 free text was represented in CSV format.

Although all the documents in Reuters-21578 had class labels associated with them in different categories, they were not explicitly organised as a hierarchy, so it was necessary to create two hierarchies (Reuters-21578-LOC-2327-2H and Reuters-21578-COM-2327-2H) based on the data set classes in order to apply the hierarchical method presented in Chapter 7. The Reuters-21578-LOC-2327-2H hierarchy comprised two levels, from the root: (i) Continent and (ii) Country; the Reuters-21578-COM-2327-2H hierarchy also comprised two levels: (i) Type of commodity and (ii) Commodity. An example fragment of the Reuters-21578 raw data is shown in Figure 3.6. Each

document has an old ID (“OLDID”) and a new ID (“NEWID”) because Reuters-21578 is an improved version of Reuters-22173, hence the dual IDs. However, neither of these IDs was used with respect to the work described in this thesis, instead a unique consecutive number was used as the ID for each document. The content of the “PLACES” tag was used to determine the two levels of the Reuters-21578-LOC-2327-2H hierarchy according to which continent(s) (first level) and to which country (second level) the document referred to. If the document referred to more than one country, the first country was considered as the most representative and the one with which the document was associated. Determining which continent a document was related to was straightforward if there was only one country to be considered. If there were more than one country, only the first two countries and their associated continents were considered: if both countries belonged to different continents a combination of the names was assigned as the class (for example “U.S.A.” and “France” would belong to the pseudo continent “AmericaEurope”), otherwise the assignment was straightforward.

Regarding the Reuters-21578-COM-2327-2H hierarchy, the content of the “TOPICS” tag was used to determine its two levels according to which type of commodity (first level) and to which specific commodity (second level) the document was related. Note that in both cases (“PLACES” and “TOPICS” tags) there could be more than one element or in some cases no element at all. In the case of many elements within a tag, the first one was chosen as it was considered to be the most representative. When there were no elements within a tag the document was ignored. The content of the “TITLE”, “DATELINE” and “BODY” tags were used together to make up the free text.

As in the case of other data sets considered in this chapter, the Reuters-21578 free text was again preprocessed according to the general free text preprocessing steps suggested by Miner *et al.* [74], as follows:

1. **Scope of documents selection:** The entire collection of free text “documents” was selected (as in the case of the SAVSNET and OHSUMED data).
2. **Case normalisation:** All alphabetic characters were again converted to lower case.
3. **Numbers, symbols and stop words removal:** Numbers and symbols and stop words were removed from the free text. As in the case of the OHSUMED data there were no particular alpha-numeric combinations that were considered significant.
4. **Tokenization:** The text was tokenized into single words in the same manner as adopted for SAVSNET and OHSUMED.

5. **Stemming:** Stemming was used to reduce the dimensionality of the data in the same manner as for the OHSUMED data.
6. **Feature selection:** Chi-squared and CFS were again used for feature selection.

As before, a variation of the above preprocessing was conducted with respect to the semi-automated method presented in Chapter 6 so that numbers and symbols were removed but phrase delimiters (commas, semicolons and full stops) were kept.

Figure 3.7 gives an example fragment of partially preprocessed Reuters-21578 records (all lower case; numbers, symbols and stop words removed; and bag-of-words format). Note the codes used to identify the class labels (bold font) and the level in the hierarchy to which each document was related (e.g. AA - First level, first class label, BC - Second level, third class label, and so on). For example, in Figure 3.7 the record with the ID “275” has the code “AC” as the class label, which corresponds to the class “Livestock”. In a similar manner, record “1745” has the code “AB” as the class label, which in turn corresponds to the class “Grains”. A similar class label coding was used with respect to the OHSUMED data sets. As in the case of the SAVSNET-917-4H data, the same documents appear in each level of the hierarchy but with different class labels in each case.

3.6 Summary

From the foregoing, the data sets used in the research described in this thesis all comprised real-world data. The main difference between SAVSNET and the other data sets was that the SAVSNET data represented a genuine questionnaire data set. A second distinction was that the SAVSNET data set contained a small amount of data and is of restricted use; while the OHSUMED and Reuters-21578 data sets contained a larger amount of data, are readily available on-line and are frequently considered as benchmark data sets in the literature. The relevance of using SAVSNET-840-4 and SAVSNET-971-3 was that they allowed investigation of the effect of using parts of the tabular element of questionnaire data in the text summarisation process by using extended versions of them (SAVSNET-840-4-TD+FT and SAVSNET-971-3-TD+FT) where tabular attributes were added to the free text in the form of words. Table 3.13 lists the complete collection of data sets used to evaluate the work described in this thesis together with an indication of the proposed methods where they were used. As mentioned in Chapter 1, the summarisation methods based on text classification considered in this thesis are: (i) Standard Classification (SC), (ii) Classifier Generation Using Secondary Data (CGUSD), (iii) Semi-Automated Rule Summarisation Extraction Tool (SARSET) and (iv) Hierarchical Text Classification (HTC). From Table 3.13 it can be noted that all the data sets were used in all cases with the exception of SAVSNET-840-

```

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="18913"
NEWID="2495">
<DATE> 5-MAR-1987 17:10:15.09</DATE>
<TOPICS><D>livestock</D><D>carcass</D></TOPICS>
<PLACES><D>usa</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN>
&#5;&#5;&#5;F
&#22;&#22;&#1;f0943&#31;reute
r f BC-MEATPACKERS-REJECT-OC 03-05 0101</UNKNOWN>
<TEXT>&#2;
<TITLE>MEATPACKERS REJECT OCCIDENTAL &lt;OXY> UNIT OFFER</TITLE>
<DATELINE> CHICAGO, March 5 - </DATELINE><BODY>United Food and Commercial
Workers Union Local 222 rejected a new contract proposal from Iowa Beef Processors Inc and
remain out of work, union spokesman Allen Zack said. In mid-December, Iowa Beef, a subsidiary
of Occidental Petroleum Corp, closed its beef processing plant at Dakota City, Nebraska, because
it said "it had no alternative to threats by meatpackers to disrupt operations." About 2,800 UFCWU
members are affected by what the union terms as a lockout. A 3-1/2 year labor contract at the plant
expired December 13. Zack said IBP's proposal included elimination of a two-tier wage structure,
a 60 cent an hour wage cut for slaughterers and a 45 cent an hour wage reduction for processors.
The new proposal also included a bonus system of 1,000 dlrs for workers who had been at the
plant for two years, Zack said. The annual turnover rate at the facility is 100 pct, he said. Reuter
&#3;</BODY></TEXT>
</REUTERS>

<REUTERS TOPICS="YES" LEWISSPLIT="TEST" CGISPLIT="TRAINING-SET" OLDID="682"
NEWID="16147">
<DATE>13-APR-1987 05:42:03.66</DATE>
<TOPICS><D>grain</D><D>corn</D><D>rice</D><D>oilseed</D><D>cottonseed</D>
<D>groundnut</D></TOPICS>
<PLACES><D>china</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN>
&#5;&#5;&#5;G C
&#22;&#22;&#1;f0331&#31;reute
u f BC-CHINA-RAISES-GRAIN-PU 04-13 0099</UNKNOWN>
<TEXT>&#2;
<TITLE>CHINA RAISES GRAIN PURCHASE PRICES</TITLE>
<DATELINE> PEKING, April 13 - </DATELINE><BODY>China has raised the state purchase
prices of corn, rice, cottonseed and shelled peanuts from April 1 to encourage farmers to grow
them, the official China Commercial Daily said. The paper said the price paid for corn from 14
northern provinces, cities and regions has increased by one yuan per 50 kg. A foreign agricultural
expert said the rise will take the price to 17 fen per jin (0.5 kg) from 16 fen. The paper said the price
for long-grained rice from 10 southern provinces and cities was raised by 1.5 yuan per 50 kg. The
paper said the price for round-grained rice from 11 provinces, regions and cities in central, east
and northwest China has been increased by 1.75 yuan per 50 kg. It gave no more price details. It
said local authorities must inform farmers of the price increases before farmers begin planting, to
encourage production of grains and oilseeds. Chinese officials have said farmers are unwilling to
grow grain because they can earn more from other crops. REUTER &#3;</BODY></TEXT>
</REUTERS>

```

Figure 3.6: Example fragment of raw Reuters-21578 data.

id	freeText	class
275	meatpackers reject occidental oxy unit offer chicago march united food commercial workers union local rejected contract proposal iowa beef processors remain work union spokesman allen zack mid december iowa beef subsidiary occidental petroleum corp closed beef processing plant dakota city nebraska alternative eats meatpackers disrupt operations ufcwu members union terms lockout year labor contract plant expired december zack ibp proposal included elimination tier wage structure cent hour wage cut slaughterers cent hour wage reduction processors proposal included bonus system dlrs workers plant years zack annual turnover rate facility pct reuter,	AC
1745	china raises grain purchase prices peking april china raised purchase prices corn rice cottonseed shelled peanuts april encourage farmers grow official china commercial paper price paid corn northern provinces cities regions increased yuan foreign agricultural expert rise will price fen jin fen paper price long grained rice southern provinces cities raised yuan paper price round grained rice provinces regions cities central east northwest china increased yuan price details local authorities inform farmers price increases farmers planting encourage production grains oilseeds chinese officials farmers unwilling grow grain earn crops reuter,	AB

Figure 3.7: Two example preprocessed Reuters-21578 records.

4-FT, SAVSNET-840-4-TD+FT, SAVSNET-971-3-FT and SAVSNET-971-3-TD+FT, which were not used in the HTC experiments as they did not feature class hierarchies.

Thus, in summary, this chapter has described the data sets used to evaluate the text summarisation techniques proposed in this thesis as well as describing the respective preprocessing applied in each case. It was noted that the tabular and the free text elements of questionnaire data, having different characteristics, were preprocessed separately. The preprocessing steps in each case were described in detail. The idea of merging relevant tabular data attributes with free text as a means of enriching the textual element of a questionnaire data was also introduced.

Table 3.13: Comparison of data sets and their usage in relation to the proposed methods.

Data sets			Text Classification methods for Text Summarisation			
Name	Classes	Docs.	SC	CGUSD	SARSET	HTC
SAVSNET-840-4-FT	4	840	✓	✓	✓	✗
SAVSNET-840-4-TD+FT	4	840	✓	✗	✗	✗
SAVSNET-971-3-FT	3	971	✓	✓	✓	✗
SAVSNET-971-3-TD+FT	3	971	✓	✗	✗	✗
SAVSNET-917-1L	3	917	✓	✓	✓	✓
SAVSNET-917-2L	3	917	✓	✗	✓	✓
SAVSNET-917-3L	3	917	✓	✗	✓	✓
SAVSNET-917-4L	5	917	✓	✗	✓	✓
OHSUMED-CA-3187-1L	2	3,187	✓	✓	✓	✓
OHSUMED-CA-2570-2L	16	2,570	✓	✓	✓	✓
OHSUMED-CA-834-3L	7	834	✓	✓	✓	✓
OHSUMED-AD-3393-1L	34	3,393	✓	✓	✓	✓
OHSUMED-AD-569-2L	26	569	✓	✓	✓	✓
OHSUMED-AD-292-3L	9	292	✓	✓	✓	✓
Reuters-21578-LOC-2327-1L	14	2,327	✓	✓	✓	✓
Reuters-21578-LOC-2327-2L	92	2,327	✓	✓	✓	✓
Reuters-21578-COM-2327-1L	5	2,327	✓	✓	✓	✓
Reuters-21578-COM-2327-2L	52	2,327	✓	✗	✓	✓

Chapter 4

Using Standard Classification Techniques for Text Summarisation

4.1 Introduction

The previous chapter presented how the questionnaire data used for evaluation purposes with respect to this thesis was preprocessed and represented prior to applying the text summarisation techniques proposed later in this thesis. The proposed techniques are considered in this and in the following three chapters. As noted in Chapter 2, a number of approaches are reported in the literature that are directed at the summarisation of text documents using text classification methods [9, 89, 58, 43], although these approaches operate in a different manner to that presented in this thesis. This chapter considers the potential of using a single standard classification technique to generate the desired summarisation classifiers. More specifically the work described in this chapter had four objectives:

1. To establish a benchmark with which to compare the alternative summarisation classification techniques presented later in this thesis.
2. To compare the operation of a number of different classifiers on data sets containing different types of text and select the ones that produced the best performance for use with respect to the work described later in Chapters 5 and 7.
3. To ascertain whether the addition of tabular data was useful or not in terms of the quality of the classification results and consequently in the quality of the summaries.
4. To compare the operation of two potential feature selection techniques and determine which one produced the best performance.

It was therefore considered desirable to compare and identify which classification technique produced the best performance with respect to the different kinds of data sets considered (questionnaires, academic papers and news articles). The standard classification techniques and the evaluation measures used were described extensively in Chapter 2.

The remainder of this chapter is arranged as follows. The proposed summarisation methodology is described in Section 4.2. Section 4.3 presents a comprehensive description of the experiments carried out on the data sets, as well as an interpretation of the obtained results, in the context of the proposed methodology. A summary of the evaluation results obtained is presented in Section 4.4 and some discussion of the results in Section 4.5. Finally, a summary of the chapter is presented in Section 4.6.

4.2 Methodology

A schematic illustration of the summarisation approach using standard classification techniques is presented in Figure 4.1. From the figure, the input consists of data sets containing text documents which may also include tabular data (as in the case of questionnaire data). The approach comprises four main stages: (i) preprocessing and representation of the input data, (ii) feature selection, (iii) classification and (iv) summarisation. The first stage (shown in Figure 4.1 in the area labelled “(1) Pre-processing”) was comprehensively described in a generic manner in Chapter 3. As explained in Chapter 3, two feature selection techniques were used, (i) TF-IDF in conjunction with Chi-squared and (ii) TF-IDF in conjunction with CFS, and thus these are not discussed further here. The last two stages are the most significant with respect to the approach presented in this chapter and are described in detail in the following two subsections.

4.2.1 Classification

As shown in the area labelled “(3) Text Classification” in Figure 4.1 a classifier is applied to the input data. How this classifier is best generated was one of the objectives of the work described in this chapter. Experiments were conducted (reported later in this chapter) using seven different standard classification techniques as follows:

- TFPC
- C4.5
- SMO
- LibSVM
- Naive Bayes
- K-Nearest Neighbour
- RIPPER

These were selected because: (i) they are well known classification techniques whose

usage has been widely reported in the literature and (ii) they display a variety of methods of operation. Each of the adopted classifier generators was introduced in Chapter 2. Once the classifiers have been applied to the data sets, they may be evaluated. With respect to the work described in this thesis the following two step classifier generation procedure was adopted:

1. For each classification technique in question, train using stratified Ten-fold Cross Validation (TCV).
2. For each generated classifier evaluate using five evaluation measures: (i) overall accuracy expressed as a percentage, (ii) Area Under the ROC Curve (AUC), (iii) precision, (iv) sensitivity/recall and (v) specificity. Of these, accuracy and AUC were considered to be the most relevant; precision, sensitivity/recall and specificity are presented so as to provide a broader insight into the effectiveness of the individual classifiers. (Each of these measures was described in Chapter 2).

The classifier generation process that produced the best classification results was then further used to evaluate the summarisation techniques described in Chapters 5 to 7.

4.2.2 Summary Generation

As shown in Figure 4.1 in the area labelled “(4) Text Summarisation”, the classes assigned to unseen documents as a result of applying a selected classifier were used to generate the desired summaries by prepending or appending “canned” text to individual class labels, and in consequence increasing the readability of the extracted meaning. Simple rules of the form:

$$\text{if } \langle CLASS_NAME \rangle \text{ then } \langle PREPEND_APPEND \rangle \\ \langle DOMAIN_SPECIFIC_TEXT \rangle$$

were established by recourse to domain experts. These rules were then used to prepend or append domain-specific text to generated class labels (names). Some example rules are presented in Table 4.1 where the last rule is treated as a “catch all” default rule. Note that, with respect to Table 4.1, in some cases, as in the case of class1 and class3, the same text may be appended or prepended.

Algorithm 1 shows how the names of the classes that are assigned to text documents, after the application of a classifier, are coupled with domain-specific text according to rules of the above form. The input to the algorithm is: (i) a class label c_i associated with a particular document and (ii) a set of n rules $R = \{r_1, r_2, \dots, r_n\}$. Domain-specific text is prepended/appended to the name of the class according to the rules in R . In the rare case where no class was assigned, default domain-specific text is used as

Table 4.1: Example of rules for generating summaries.

if (class1) then (prepend) (''This document was about'')
if (class2) then (append) (''was the main topic of this document'')
if (class3) then (prepend) (''This document was about'')
if (class4) then (append) (''was the main topic of this document'')
if (noClass) then (append) (''This document was about <domain area>'')

defined by rule r_n . The output is a text summarisation which is enclosed by quotation marks.

Data: $action, c_i, R = \{r_1, r_2, \dots, r_n\}$
Result: text summary of the original document which includes c_i and prepended/appended text

```

1  $s = null\_string$ 
2 for all  $r_i$  in  $R$  from  $i = 1$  to  $i = n - 1$  do
3   | if  $r_i.ancestor = c_i$  then
4   |   | if  $action = prepend$  then
5   |   |   |  $s \leftarrow c_i$  with  $r_i.consequent$  prepended
6   |   | else if  $action = append$  then
7   |   |   |  $s \leftarrow c_i$  with  $r_i.consequent$  appended
8 end
9 if  $s = null\_string$  then
10  | if  $action = prepend$  then
11  |   |  $s \leftarrow c_i$  with  $r_i.consequent$  prepended
12  | else if  $action = append$  then
13  |   |  $s \leftarrow c_i$  with  $r_i.consequent$  appended
14 return  $s$ 
```

Algorithm 1: Prepending/appending text to a name of a class.

As was previously highlighted in Chapter 1, in this thesis text summarisation is conceived as a form of text classification in that the classes assigned to text documents are viewed as an indication of the main ideas of the original free text but in a coherent and reduced form. Therefore, it is suggested that summaries of the forms: (i) *Domain-specific text + name of class* and (ii) *Name of class + domain-specific text*, are relevant, especially where the text is unstructured and contains features such as the ones found in the free text part of questionnaires (misspelled words, poor grammar, and abbreviations and acronyms related to a specific domain).

4.3 Experiments and Results

This section reports on the experiments conducted to evaluate the use of standard classification techniques for summary generation. As noted above, seven standard classification techniques were applied to the 18 identified data sets (8 of which contained

questionnaire data). Recall that during preprocessing two different feature selection methods were considered, Chi-squared and CFS, thus giving a total of 36 variations (the two feature selection methods applied to each of the 18 data sets).

Note that TFPC was applied using four different confidence threshold (γ) values: (i) 50%, (ii) 60%, (iii) 70% and (iv) 80%. These four variations of TFPC are identified below as TFPC-C50, TFPC-C60, TFPC-C70 and TFPC-C80. Thus ten standard classification techniques were considered in total. Note also that in the case of the TFPC algorithm a range of support threshold (σ) values, from 0.5 to 2.5, incremented in steps of 0.5, were used. The results presented later in this section are therefore averages over the range of (σ) values with respect to each (γ) value.

The classification results for each data set are presented in Tables 4.2 to 4.19. In the tables, the first column lists the classification technique used: (i) TFPC-C50, (ii) TFPC-C60, (iii) TFPC-C70, (iv) TFPC-C80, (v) C4.5, (vi) SMO, (vii) LibSVM, (viii) NB (Naive Bayes), (ix) KNN (K-Nearest-Neighbour) and (x) RIPPER. The remaining ten columns are divided into two groups. From Table 4.2 to Table 4.5 the first five columns correspond to the free text only and the remaining five to the free text combined with tabular data. From Table 4.6 to Table 4.19 the remaining ten columns after the “Method” column are also divided into two groups: one for each feature selection method used. For all the tables each five-column group comprises a listing of the following: (i) overall accuracy expressed as a percentage (Acc), (ii) Area Under the ROC Curve (AUC), (iii) precision (Pr), (iv) sensitivity/recall (Sn/Re) and (v) specificity (Sp). For each data set used, and taking into account the feature selection method applied, the highest recorded accuracy and AUC values are indicated in bold font. In some cases a particular classification method produced both the highest accuracy and AUC value, while in other cases one method produced the highest accuracy and another the highest AUC. The classification methods that achieved the best classification results were considered to be the ones most appropriate to text summarisation.

An overall comparison of the best classification techniques according to the recorded results is presented in Table 4.20. The rest of this section is divided into subsections each directed at one of the data sets considered and each ends with a summary of the evaluation results.

4.3.1 SAVSNET-840-4

The results presented in Tables 4.2 and 4.3 correspond to the experiments carried out using the SAVSNET-840-4-FT and the SAVSNET-840-4-TD+FT data sets respectively. Comparing the results obtained using Chi-squared feature selection (Table 4.2), it is not clear whether there is any advantage of including tabular data with the free text since many of the results are similar. The highest accuracy and AUC values in each case were obtained using SMO. In SAVSNET-840-4-FT the second highest accuracy

value (88.33%) was obtained using C4.5 while the other highest AUC value (0.95) was obtained using RIPPER. In the case of SAVSNET-840-4-TD+FT the highest accuracy and AUC values were obtained using SMO and RIPPER respectively, while SMO obtained an accuracy of 90.60% and an AUC of 0.96, RIPPER obtained an accuracy of 89.40% of accuracy and an AUC of 0.96.

From the comparison between the SAVSNET-840-4-FT and the SAVSNET-840-4-TD+FT data sets using Chi-squared (see Table 4.2), it is noticeable that for the ten classification methods, the methods that performed better were evenly distributed between the case when they were applied to the free text only and the case when they were applied to the free text combined with tabular data. The highest results overall, however, were achieved when tabular data was used in conjunction with free text. The classification methods that had the best performance were: (i) C4.5, (ii) SMO and (iii) RIPPER.

Considering the classification results obtained using CFS feature selection (Table 4.3) with respect to SAVSNET-840-4-FT and SAVSNET-840-4-TD+FT, it is interesting to note that the best classification results were achieved when considering only the free text. This contradicts the results reported in Table 4.2, the probable cause appears to be the feature selection method applied, leading to the conjecture that for the SAVSNET-840-4 data set the best results are obtained when: (i) Chi-squared feature selection is applied to free text combined with tabular data and (ii) CFS feature selection is applied to free text on its own. Interestingly most of the classification techniques performed better when only the free text SAVSNET data was considered (SAVSNET-840-4-FT). Again, the three best classification techniques were: (i) C4.5, (ii) SMO and (iii) RIPPER.

Overall, comparing the results presented in both Table 4.2 and 4.3 it can be conjectured that for generating high quality summaries from the SAVSNET-840-4 data the inclusion of additional tabular data and Chi-squared feature selection can enhance the result. The classification methods that achieved the best results in general were: (i) C4.5, (ii) SMO and (iii) RIPPER, with SMO producing the best overall performance. It is interesting to note that these classification methods all operate in a different manner.

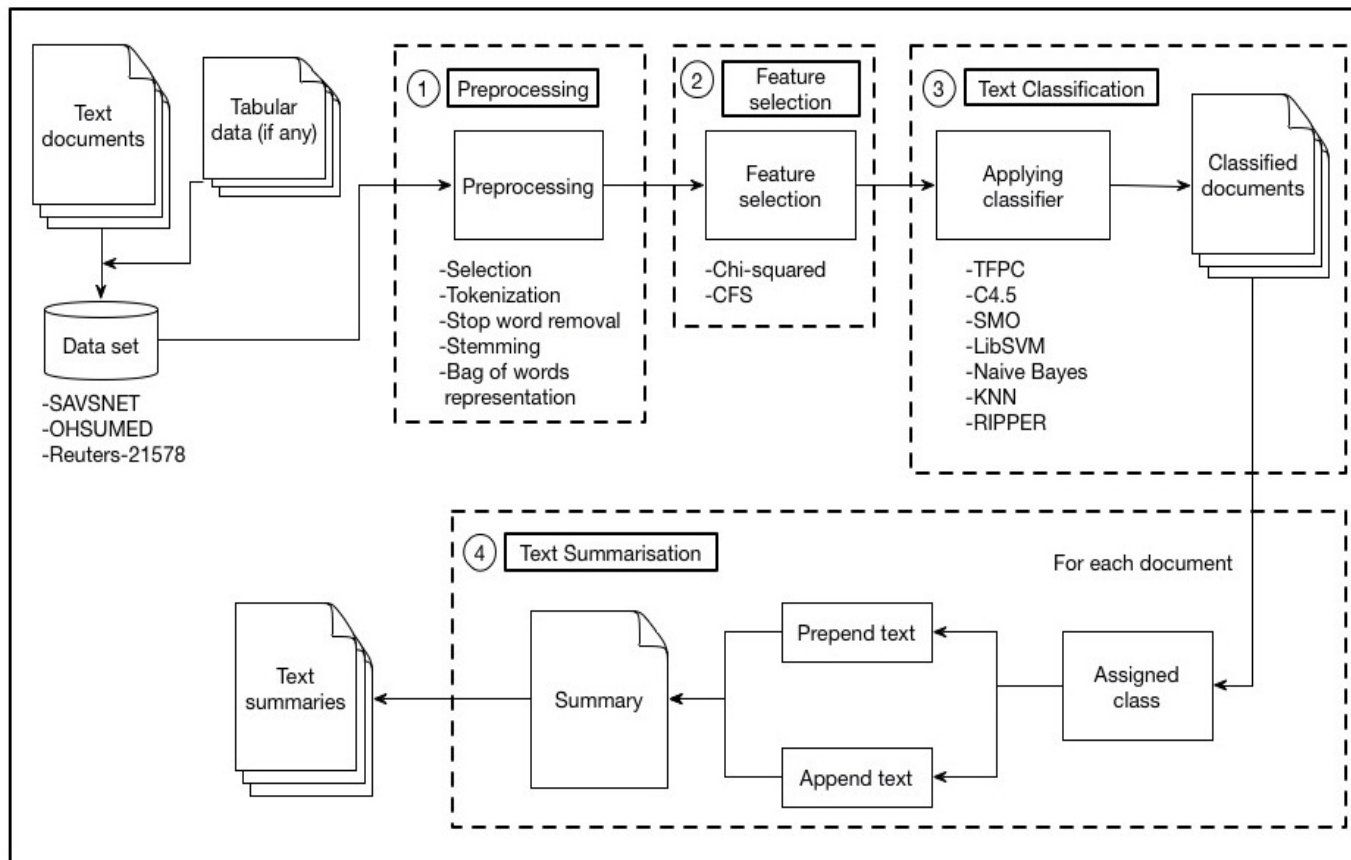


Figure 4.1: Standard Classification Technique for Text Summarisation.

Table 4.2: Classification results for the SAVSNET-840-4 data set with Chi-squared.

	Free text					Free text with tabular data				
Method	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
TFPC-C50	80.12	0.55	0.61	0.55	0.92	79.14	0.54	0.60	0.54	0.91
TFPC-C60	79.67	0.55	0.60	0.55	0.92	78.72	0.54	0.59	0.54	0.92
TFPC-C70	80.24	0.56	0.60	0.56	0.92	79.31	0.55	0.58	0.54	0.92
TFPC-C80	79.74	0.55	0.55	0.55	0.92	79.19	0.54	0.56	0.54	0.91
C4.5	88.33	0.94	0.88	0.88	0.95	88.57	0.94	0.89	0.89	0.95
SMO	89.29	0.95	0.90	0.89	0.96	90.60	0.96	0.91	0.91	0.96
LibSVM	78.33	0.83	0.75	0.78	0.88	77.02	0.82	0.74	0.77	0.87
NB	83.33	0.93	0.83	0.83	0.93	83.45	0.94	0.83	0.84	0.93
KNN	45.83	0.68	0.69	0.46	0.89	65.60	0.78	0.79	0.66	0.91
RIPPER	87.50	0.95	0.88	0.88	0.95	89.40	0.96	0.90	0.89	0.96

Table 4.3: Classification results for the SAVSNET-840-4 data set with CFS.

	Free text					Free text with tabular data				
Method	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
TFPC-C50	77.09	0.52	0.56	0.54	0.91	70.54	0.46	0.47	0.46	0.88
TFPC-C60	77.33	0.52	0.57	0.54	0.91	71.19	0.47	0.44	0.47	0.89
TFPC-C70	78.16	0.53	0.56	0.55	0.91	72.60	0.48	0.44	0.48	0.89
TFPC-C80	75.80	0.51	0.53	0.53	0.91	69.04	0.45	0.40	0.45	0.88
C4.5	88.33	0.94	0.88	0.88	0.95	85.12	0.94	0.85	0.85	0.94
SMO	86.43	0.94	0.86	0.86	0.94	83.93	0.92	0.83	0.84	0.92
LibSVM	76.19	0.82	0.72	0.76	0.87	76.64	0.80	0.71	0.75	0.85
NB	76.31	0.93	0.78	0.76	0.92	64.76	0.88	0.73	0.65	0.90
KNN	71.43	0.79	0.69	0.71	0.86	75.95	0.83	0.74	0.76	0.89
RIPPER	86.19	0.95	0.86	0.86	0.94	85.33	0.94	0.85	0.86	0.93

4.3.2 SAVSNET-971-3

Tables 4.4 and 4.5 show the results obtained with respect to the SAVSNET-971-3-FT and SAVSNET-971-3-TD+FT data sets. These tables are structured in a similar manner to Tables 4.2 and 4.3.

Considering the results obtained using Chi-squared feature selection first (Table 4.4) it is noticeable that, although the free text only data has produced consistently better accuracy and AUC, the best results were the ones obtained when SMO was applied to free text combined with tabular data (recorded accuracy of 75.30% and AUC of 0.80). However, the results for the free text only data set using SMO are very similar, having an accuracy of 75.87% and an AUC of 0.80.

In Table 4.5 (CFS feature selection) most of the results obtained using free text only were higher than the ones obtained using free text and tabular data in conjunction. Similarly to the results obtained using Chi-squared feature selection, the best accuracy and AUC results using CFS were obtained with SMO (76.21% and 0.79 respectively). In the case of free text combined with tabular data the best accuracy value (70%) was achieved using TFPC-C60 and the best AUC value (0.75) using C4.5.

Overall, in the context of SAVSNET-971-3-FT, most of the best results were obtained using Chi-squared feature selection. However, the two best results achieved (using SMO and C4.5) with respect to SAVSNET-971-3-TD+FT, were slightly higher than the ones achieved using SAVSNET-971-3-FT. RIPPER was the third best method in both cases. When CFS was used to select features most of the best results were obtained with respect to SAVSNET-971-3-FT. In the case of SAVSNET-971-3-TD+FT, TFPC-C60 and SMO produced the best accuracy values, however, considering also AUC, the two best results were once again obtained using C4.5 and with SMO.

Table 4.4: Classification results for the SAVSNET-971-3 data set with Chi-squared.

	Free text					Free text with tabular data				
Method	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
TFPC-C50	69.46	0.46	0.53	0.46	0.76	67.98	0.44	0.54	0.44	0.74
TFPC-C60	69.71	0.46	0.54	0.46	0.76	68.82	0.46	0.55	0.46	0.76
TFPC-C70	69.93	0.48	0.52	0.48	0.77	69.45	0.47	0.52	0.47	0.77
TFPC-C80	67.40	0.45	0.46	0.45	0.75	67.10	0.45	0.50	0.45	0.75
C4.5	70.34	0.75	0.69	0.70	0.76	70.53	0.76	0.69	0.71	0.77
SMO	74.87	0.80	0.74	0.75	0.79	75.30	0.80	0.75	0.75	0.81
LibSVM	56.75	0.51	0.50	0.57	0.45	56.10	0.50	0.48	0.56	0.45
NB	60.25	0.74	0.65	0.60	0.75	59.86	0.73	0.64	0.60	0.75
KNN	59.42	0.63	0.64	0.59	0.68	51.32	0.68	0.73	0.51	0.85
RIPPER	69.52	0.72	0.66	0.70	0.71	68.29	0.70	0.65	0.65	0.67

Table 4.5: Classification results for the SAVSNET-971-3 data set with CFS.

	Free text					Free text with tabular data				
Method	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
TFPC-C50	67.07	0.43	0.50	0.43	0.74	68.40	0.45	0.48	0.45	0.76
TFPC-C60	69.34	0.47	0.50	0.47	0.77	70.00	0.48	0.49	0.48	0.78
TFPC-C70	68.91	0.46	0.48	0.47	0.76	67.45	0.45	0.46	0.45	0.76
TFPC-C80	66.47	0.43	0.45	0.43	0.74	65.48	0.43	0.41	0.43	0.74
C4.5	70.13	0.76	0.68	0.70	0.75	68.70	0.75	0.66	0.69	0.70
SMO	76.21	0.79	0.75	0.76	0.78	69.00	0.70	0.65	0.69	0.67
LibSVM	57.16	0.51	0.50	0.57	0.46	61.59	0.54	0.56	0.62	0.46
NB	55.72	0.72	0.65	0.56	0.78	44.41	0.66	0.58	0.44	0.74
KNN	67.35	0.63	0.66	0.67	0.59	63.41	0.65	0.62	0.63	0.66
RIPPER	69.72	0.70	0.67	0.70	0.67	67.48	0.68	0.64	0.68	0.62

4.3.3 SAVSNET-917-4H

As was mentioned earlier in this chapter, each one of the remaining data sets used in the experiments contains only free text and were part of a hierarchy of classes. However, for the purpose of generating summaries using standard classification methods, each level of the hierarchy for each data set was considered as an independent data set. Tables 4.6, 4.7, 4.8 and 4.9 present the results for each of the data sets that comprise the SAVSNET-917-4H hierarchy of classes.

4.3.3.1 SAVSNET-917-1L

Table 4.6 shows the results obtained for the first level of the SAVSNET-917 data set, thus SAVSNET-917-1L. In the case of Chi-squared feature selection the classification methods that produced the best accuracy values were SMO and RIPPER (70.34% and 67.18% respectively). The best AUC values were obtained using SMO (0.75) and Naive Bayes (0.70). On the other hand when CFS feature selection was used, SMO and RIPPER produced the best accuracy values (67.61% and 67.07%), and SMO and C4.5 the best AUC values (0.72 and 0.75 respectively).

Regardless of the feature selection method used, SMO proved to be the best overall classification method, followed very closely by RIPPER. C4.5 did not perform well with respect to accuracy when Chi-squared feature selection was used. When CFS feature selection was used, C4.5 performed better than TFPC. In general the best results were obtained with Chi-squared feature selection, however the values obtained with both feature selection methods were similar, with SMO and RIPPER being identified as the best classification methods in both cases.

4.3.3.2 SAVSNET-917-2L

Table 4.7 shows the results for the SAVSNET-917-2L data set. When Chi-squared feature selection was used, the best accuracy values were obtained using TFPC-C70 (66.02%) and with RIPPER (66.96%). The best AUC values were obtained using SMO (0.57) and RIPPER (0.57). Similarly, when CFS feature selection was used to select features, the classification method that generated both the best accuracy and AUC values was RIPPER (68.38% and 0.58 respectively). The best AUC value was obtained using SMO (0.60). The second best accuracy value was obtained using TFPC-C50 (66.72%). The other TFPC variants also performed relatively well with respect to RIPPER in terms of accuracy. It was very evident that the best results in general were obtained using CFS feature selection. Regardless of which feature selection technique was used, the highest values were obtained using TFPC, RIPPER and SMO.

Table 4.6: Classification results for the SAVSNET-917-1L data set with Chi-squared and CFS.

Method	Chi-squared					CFS				
	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
TFPC-C50	64.27	0.42	0.49	0.42	0.73	61.20	0.38	0.43	0.38	0.70
TFPC-C60	66.08	0.45	0.48	0.45	0.74	62.31	0.40	0.47	0.40	0.71
TFPC-C70	64.66	0.42	0.47	0.42	0.73	60.61	0.38	0.41	0.38	0.70
TFPC-C80	62.91	0.41	0.45	0.41	0.72	59.54	0.35	0.33	0.35	0.68
C4.5	63.14	0.69	0.61	0.63	0.71	66.09	0.75	0.64	0.66	0.74
SMO	70.34	0.75	0.68	0.70	0.75	67.61	0.72	0.65	0.68	0.71
LibSVM	56.38	0.53	0.50	0.56	0.49	62.27	0.55	0.57	0.62	0.48
NB	58.67	0.70	0.61	0.59	0.73	49.07	0.70	0.62	0.49	0.78
KNN	33.81	0.57	0.55	0.34	0.79	53.11	0.57	0.52	0.53	0.62
RIPPER	67.18	0.69	0.63	0.67	0.67	67.07	0.66	0.64	0.67	0.65

Table 4.7: Classification results for the SAVSNET-917-2L data set with Chi-squared and CFS.

Method	Chi-squared					CFS				
	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
TFPC-C50	65.87	0.35	0.34	0.35	0.68	66.72	0.35	0.40	0.35	0.68
TFPC-C60	65.92	0.36	0.38	0.36	0.68	66.46	0.36	0.40	0.36	0.69
TFPC-C70	66.02	0.35	0.34	0.35	0.68	66.35	0.36	0.38	0.36	0.68
TFPC-C80	65.94	0.34	0.28	0.34	0.67	66.19	0.35	0.32	0.35	0.68
C4.5	56.71	0.56	0.54	0.57	0.53	59.11	0.55	0.56	0.59	0.53
SMO	57.58	0.57	0.55	0.58	0.54	62.70	0.60	0.59	0.63	0.54
LibSVM	64.45	0.51	0.54	0.64	0.38	64.89	0.51	0.54	0.65	0.37
NB	40.89	0.53	0.51	0.41	0.63	39.15	0.55	0.53	0.39	0.67
KNN	36.53	0.52	0.52	0.37	0.69	56.16	0.51	0.51	0.56	0.46
RIPPER	66.96	0.57	0.60	0.67	0.45	68.38	0.58	0.63	0.68	0.47

4.3.3.3 SAVSNET-917-3L

The results for SAVSNET-917-3L are shown in Table 4.8. Using both Chi-squared and CFS feature selection the best accuracy values were obtained using TFPC, with TFPC-C50 producing the best accuracy (63.08% using Chi-squared and 62.94% using CFS). The second best accuracy values were obtained by TFPC-C80 using Chi-squared with 62.92% and by TFPC-C60 using CFS with 62.83%. In terms of AUC, the classification methods that produced the best performance were C4.5 and SMO (0.59 and 0.58 using Chi-squared, and 0.57 and 0.58 using CFS respectively). Naive Bayes also produced a good AUC value in combination with CFS feature selection. Overall, for the SAVSNET-917-3L data set, the results obtained using CFS feature selection were slightly better than the ones obtained using Chi-squared feature selection. In both cases, while the best accuracy results were obtained by variants of TFPC, the best AUC results were obtained using C4.5 and SMO.

4.3.3.4 SAVSNET-917-4L

Table 4.9 presents the results for the SAVSNET-917-4L data set. The two highest accuracy values (obtained using Chi-squared feature selection) were 45.48% (using TFPC-C60) and 47% (using SMO). When using CFS feature selection, SMO and RIPPER produced the highest accuracy results (45.37% and 44.49% respectively). In terms of AUC, in both cases the best values were obtained using SMO and Naive Bayes. In general the results obtained were not as good as expected; the cause might have been the unbalanced nature of the data set. Overall the best results were obtained using SMO.

Table 4.8: Classification results for the SAVSNET-917-3L data set with Chi-squared and CFS.

	Chi-squared					CFS				
Method	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
TFPC-C50	63.08	0.35	0.36	0.35	0.68	62.94	0.35	0.36	0.35	0.68
TFPC-C60	62.62	0.35	0.35	0.35	0.68	62.83	0.35	0.36	0.35	0.68
TFPC-C70	62.88	0.34	0.31	0.34	0.68	62.76	0.34	0.32	0.34	0.68
TFPC-C80	62.92	0.34	0.27	0.34	0.67	62.81	0.34	0.26	0.34	0.68
C4.5	57.25	0.59	0.56	0.57	0.57	56.27	0.57	0.54	0.56	0.55
SMO	56.60	0.58	0.56	0.57	0.56	60.85	0.58	0.57	0.61	0.55
LibSVM	58.12	0.50	0.49	0.58	0.41	60.52	0.51	0.50	0.61	0.41
NB	43.95	0.55	0.54	0.44	0.64	44.82	0.57	0.56	0.45	0.67
KNN	43.62	0.56	0.58	0.44	0.68	49.07	0.54	0.53	0.49	0.59
RIPPER	62.60	0.54	0.55	0.63	0.44	62.38	0.54	0.55	0.62	0.43

Table 4.9: Classification results for the SAVSNET-917-4L data set with Chi-squared and CFS.

	Chi-squared					CFS				
Method	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
TFPC-C50	43.38	0.21	0.22	0.22	0.81	43.78	0.21	0.23	0.23	0.81
TFPC-C60	45.48	0.22	0.23	0.23	0.81	43.13	0.21	0.23	0.23	0.81
TFPC-C70	42.32	0.20	0.18	0.21	0.79	43.11	0.20	0.19	0.22	0.80
TFPC-C80	33.75	0.16	0.13	0.17	0.63	23.43	0.11	0.11	0.12	0.45
C4.5	38.39	0.55	0.37	0.38	0.69	38.06	0.54	0.36	0.38	0.67
SMO	47.00	0.64	0.47	0.47	0.75	45.37	0.59	0.42	0.45	0.69
LibSVM	39.37	0.49	0.31	0.39	0.58	42.86	0.50	0.32	0.43	0.56
NB	35.55	0.58	0.42	0.36	0.76	27.26	0.57	0.37	0.27	0.79
KNN	26.83	0.54	0.39	0.27	0.81	37.19	0.53	0.36	0.37	0.68
RIPPER	43.95	0.51	0.38	0.44	0.58	44.49	0.53	0.39	0.45	0.59

4.3.4 OHSUMED-CA-3187-3H

Tables 4.10, 4.11 and 4.12 present the results for each of the data sets that comprise the OHSUMED-CA-3187-3H hierarchy of classes.

4.3.4.1 OHSUMED-CA-3187-1L

The results for the OHSUMED-CA-3187-1L data set are presented in Table 4.10. From the table it is noticeable that almost all the accuracy and AUC values, using both the Chi-squared and the CFS feature selection methods, are very high. This might again be a consequence of having very unbalanced data where the class “Congenital Heart Defects” comprises 73.39% of the records while the class “Vascular Malformations” contains the remaining 26.61%. In cases like this, AUC provides a better insight into the operation of the classifier than accuracy.

SMO produced the best accuracy values of 94.84% using Chi-squared feature selection and 94.30% using CFS feature selection. The second best accuracy results were obtained with RIPPER using Chi-squared feature selection (93.39%) and C4.5 using CFS feature selection (92.27%). In the case of AUC, the best results were obtained using SMO and Naive Bayes, and also with C4.5 when coupled with CFS feature selection.

Overall, good results were obtained with respect to the experiments carried out on the OHSUMED-CA-3187-1L data set, the probable explanation is that the unbalanced nature of the data set might have had some influence and produced an overfitted model. The best identified classification techniques were SMO and C4.5 in terms of both the accuracy and AUC. The other classification methods that produced reasonable results were Naive Bayes and RIPPER. No best feature selection method could be identified.

4.3.4.2 OHSUMED-CA-2570-2L

Table 4.11 shows the results for the OHSUMED-CA-2570-2L data set, which is not as unbalanced as OHSUMED-CA-3187-1L and has a larger number of classes (14). Using both Chi-squared and CFS feature selection the highest accuracy values were obtained using SMO (80.82% and 74.46% respectively). Similarly, SMO produced the best AUC values with 0.94 (Chi-squared) and 0.92 (CFS). The classification method producing the second best results for both accuracy and AUC varied, 0.91 using C4.5 and RIPPER coupled with Chi-squared feature selection, and 0.89 using RIPPER coupled with CFS feature selection. Overall SMO and RIPPER proved to be the methods that obtained the best results with SMO outperforming RIPPER. From the recorded results it was also clear that using Chi-squared feature selection provided an advantage.

Table 4.10: Classification results for the OHSUMED-CA-3187-1L data set with Chi-squared and CFS.

	Chi-squared					CFS				
Method	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
TFPC-C50	90.77	0.83	0.89	0.99	0.68	90.67	0.84	0.90	0.98	0.69
TFPC-C60	90.94	0.84	0.89	0.99	0.68	90.99	0.84	0.90	0.99	0.69
TFPC-C70	91.13	0.84	0.90	0.99	0.70	90.90	0.84	0.90	0.98	0.70
TFPC-C80	92.27	0.87	0.91	0.99	0.74	91.42	0.86	0.91	0.98	0.75
C4.5	93.16	0.91	0.93	0.93	0.86	92.97	0.91	0.93	0.93	0.86
SMO	94.84	0.93	0.95	0.95	0.91	94.30	0.91	0.94	0.94	0.88
LibSVM	80.07	0.64	0.80	0.80	0.49	87.07	0.77	0.88	0.87	0.67
NB	86.34	0.94	0.89	0.86	0.91	83.05	0.92	0.87	0.83	0.88
KNN	71.74	0.65	0.73	0.72	0.59	84.03	0.79	0.84	0.84	0.70
RIPPER	93.38	0.90	0.93	0.93	0.86	92.84	0.90	0.93	0.93	0.85

Table 4.11: Classification results for the OHSUMED-CA-2570-2L data set with Chi-squared and CFS.

	Chi-squared					CFS				
Method	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
TFPC-C50	74.21	0.38	0.47	0.46	0.98	69.94	0.32	0.40	0.39	0.98
TFPC-C60	72.53	0.38	0.47	0.45	0.98	66.75	0.31	0.40	0.38	0.98
TFPC-C70	71.10	0.37	0.47	0.44	0.98	64.33	0.30	0.41	0.36	0.97
TFPC-C80	64.51	0.34	0.47	0.41	0.97	52.82	0.26	0.39	0.32	0.97
C4.5	79.13	0.91	0.80	0.79	0.96	70.98	0.87	0.71	0.71	0.95
SMO	80.82	0.94	0.81	0.81	0.96	74.46	0.92	0.74	0.75	0.95
LibSVM	80.57	0.88	0.82	0.81	0.96	51.42	0.70	0.54	0.51	0.89
NB	50.02	0.89	0.67	0.50	0.97	37.46	0.80	0.55	0.38	0.96
KNN	60.67	0.77	0.63	0.61	0.93	42.05	0.68	0.45	0.42	0.92
RIPPER	79.04	0.91	0.81	0.79	0.96	71.11	0.89	0.74	0.71	0.94

4.3.4.3 OHSUMED-CA-834-3L

The results for the OHSUMED-CA-834-3L data set are shown in Table 4.12. As in the case of the experiments carried out with the OHSUMED-CA-2570-2L data set, the classification methods that produced the best overall performance were SMO and RIPPER. When the text was preprocessed using Chi-squared feature selection the top two accuracy values were 78.42% using SMO and 77.04% using RIPPER, the top two AUC values were 0.87 using C4.5 and 0.90 using SMO. In the case of CFS feature selection the best accuracy values were 76.29% using SMO and 75.78% using RIPPER. The best AUC values in this case were 0.87 (C4.5) and 0.89 (SMO). In general, slightly better results were obtained using Chi-squared feature selection. The classification technique that achieved the best accuracy and AUC results was SMO, followed by RIPPER and C4.5.

Table 4.12: Classification results for the OHSUMED-CA-834-3L data set with Chi-squared and CFS.

Method	Chi-squared					CFS				
	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
TFPC-C50	72.42	0.38	0.42	0.40	0.94	71.26	0.39	0.43	0.41	0.94
TFPC-C60	71.94	0.37	0.41	0.39	0.94	70.46	0.38	0.42	0.40	0.94
TFPC-C70	70.89	0.37	0.40	0.38	0.94	69.73	0.38	0.42	0.39	0.94
TFPC-C80	71.74	0.37	0.41	0.38	0.94	70.61	0.38	0.42	0.39	0.94
C4.5	74.40	0.87	0.75	0.74	0.92	71.89	0.87	0.72	0.72	0.92
SMO	78.42	0.90	0.79	0.78	0.92	76.29	0.89	0.75	0.76	0.92
LibSVM	46.68	0.59	0.43	0.47	0.71	62.23	0.71	0.60	0.62	0.80
NB	56.46	0.86	0.69	0.57	0.93	53.45	0.84	0.67	0.54	0.92
KNN	33.12	0.60	0.63	0.33	0.87	53.45	0.66	0.54	0.54	0.80
RIPPER	77.04	0.86	0.78	0.77	0.90	75.78	0.85	0.77	0.76	0.89

Table 4.13: Classification results for the OHSUMED-AD-3393-1L data set with Chi-squared and CFS.

Method	Chi-squared					CFS				
	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
TFPC-C50	69.23	0.04	0.09	0.07	0.97	73.71	0.02	0.04	0.04	0.97
TFPC-C60	69.19	0.04	0.09	0.07	0.97	74.99	0.02	0.04	0.04	0.97
TFPC-C70	69.27	0.04	0.09	0.07	0.97	74.99	0.02	0.04	0.04	0.97
TFPC-C80	68.80	0.04	0.08	0.07	0.97	74.99	0.02	0.04	0.04	0.97
C4.5	79.89	0.81	0.77	0.80	0.79	65.95	0.51	0.48	0.66	0.39
SMO	82.46	0.85	0.81	0.83	0.77	65.42	0.52	0.48	0.65	0.37
LibSVM	75.80	0.67	0.70	0.76	0.58	65.38	0.52	0.46	0.65	0.38
NB	40.37	0.84	0.72	0.40	0.96	38.90	0.76	0.68	0.39	0.94
KNN	55.22	0.70	0.69	0.55	0.83	46.69	0.60	0.57	0.47	0.74
RIPPER	80.71	0.80	0.79	0.81	0.78	74.17	0.71	0.71	0.74	0.62

4.3.5 OHSUMED-AD-3393-3H

Tables 4.13, 4.14 and 4.15 show the results obtained with respect to the three levels of the OHSUMED-AD-3393-3H hierarchy of classes, which were used as independent data sets.

4.3.5.1 OHSUMED-AD-3393-1L

The results for the OHSUMED-AD-3393-1L data set, which is very unbalanced and has many classes (34), are shown in Table 4.13. In general, with the exception of the results obtained using the TFPC variants, the accuracy results obtained using Chi-squared feature selection were better than the ones obtained using CFS feature selection. On the other hand, all the AUC values obtained using Chi-squared feature selection were higher than the ones obtained using CFS. The highest accuracy values obtained using Chi-squared feature selection were 82.46% (SMO) and 80.71% (RIPPER); and the best AUC values were 0.85 (SMO) and 0.84 (Naive Bayes). Using CFS feature selection, the top two accuracy values obtained were 74.99% using TFPC-C60, TFPC-C70 and TFPC-C80, and 74.17% using RIPPER. The best AUC values were 0.76 (Naive Bayes) and 0.71 (RIPPER).

Overall, the best classification method when using Chi-squared feature selection was C4.5, while in the case of CFS feature selection it was RIPPER. Note that all the classification techniques had a better performance when Chi-squared feature selection was used with the exception of the TFPC variants; which, along with RIPPER, produced the best accuracy results when CFS feature selection was used. In terms of AUC, the best performance using CFS was achieved using Naive Bayes and RIPPER. The classification techniques that produced the best results using Chi-squared feature selection were C4.5, SMO and RIPPER.

4.3.5.2 OHSUMED-AD-569-2L

Table 4.14 shows the results for the OHSUMED-AD-569-2L data set, another very unbalanced data set with many classes (26). The results, however, were considered to be acceptable since they did not seem to suggest a problem of overfitting considering the nature of the data set. When Chi-squared feature selection was used, the best accuracy values were obtained using C4.5 (76.74%) and RIPPER (72.99%), and the best AUC value (0.87) using C4.5 and Naive Bayes. With respect to CFS feature selection, on the other hand, the methods that produced the best accuracy were the ones that also featured the best AUC: an accuracy of 74.78% and AUC of 0.86 using C4.5, and an accuracy of 73.17% and AUC of 0.93 using RIPPER. Overall, the best results for both accuracy and AUC were the ones obtained using Chi-squared feature selection. The classification methods that produced the best performance were C4.5 and RIPPER.

4.3.5.3 OHSUMED-AD-292-3L

Table 4.15 presents the results for the OHSUMED-AD-292-3L data set, which consists of 9 unevenly distributed classes. The difference between the feature selection methods used is more apparent here than in any of the other cases considered, with Chi-squared feature selection producing the best results with respect to all the classification methods considered except LibSVM which worked better with CFS feature selection and KNN which produced the same accuracy and better AUC. In terms of accuracy the best classification methods (coupled with Chi-squared feature selection) were C4.5 and TFPC-C60, whereas in terms of AUC the best methods were C4.5 and Naive Bayes. For CFS feature selection, on the other hand, in terms of accuracy the best methods were C4.5 and RIPPER; in terms of AUC the best ones were C4.5 and Naive Bayes. Overall, the best classification method was C4.5. It can be conjectured that these results might be the consequence of the feature selection methods choosing different features in each case, which is something that may be expected to a certain degree, but not to the point of having such a broad difference between the results.

Table 4.14: Classification results for the OHSUMED-AD-569-2L data set with Chi-squared and CFS.

Method	Chi-squared					CFS				
	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
TFPC-C50	72.54	0.11	0.22	0.23	0.99	69.11	0.09	0.21	0.21	0.99
TFPC-C60	71.29	0.11	0.22	0.23	0.99	68.63	0.09	0.21	0.20	0.98
TFPC-C70	70.10	0.11	0.22	0.22	0.99	68.19	0.09	0.21	0.20	0.98
TFPC-C80	69.07	0.10	0.22	0.22	0.99	66.93	0.09	0.21	0.20	0.98
C4.5	76.74	0.87	0.78	0.77	0.93	74.78	0.86	0.74	0.75	0.92
SMO	67.80	0.85	0.68	0.68	0.88	71.56	0.85	0.71	0.72	0.88
LibSVM	55.46	0.66	0.53	0.56	0.77	47.23	0.60	0.55	0.47	0.73
NB	55.46	0.87	0.66	0.56	0.95	54.38	0.85	0.69	0.54	0.95
KNN	42.58	0.69	0.53	0.43	0.87	42.58	0.65	0.45	0.43	0.82
RIPPER	72.99	0.84	0.74	0.73	0.88	73.17	0.93	0.82	0.77	0.97

Table 4.15: Classification results for the OHSUMED-AD-292-3L data set with Chi-squared and CFS.

Method	Chi-squared					CFS				
	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
TFPC-C50	88.73	0.48	0.59	0.62	0.99	57.82	0.27	0.36	0.38	0.94
TFPC-C60	89.08	0.49	0.59	0.63	0.99	60.42	0.27	0.37	0.39	0.95
TFPC-C70	88.30	0.48	0.59	0.62	0.99	55.97	0.26	0.35	0.37	0.94
TFPC-C80	88.02	0.48	0.59	0.62	0.98	51.94	0.26	0.35	0.37	0.94
C4.5	89.69	0.91	0.90	0.90	0.93	69.07	0.82	0.68	0.69	0.80
SMO	82.47	0.90	0.85	0.83	0.87	67.35	0.76	0.68	0.67	0.76
LibSVM	48.11	0.53	0.38	0.48	0.58	57.39	0.61	0.64	0.57	0.65
NB	79.73	0.93	0.83	0.80	0.96	41.58	0.80	0.66	0.42	0.95
KNN	55.33	0.62	0.56	0.55	0.69	55.33	0.69	0.53	0.55	0.76
RIPPER	85.91	0.89	0.88	0.86	0.89	68.38	0.78	0.72	0.68	0.76

4.3.6 Reuters-21578-LOC-2327-2H

Tables 4.16 and 4.17 present the results of the two levels of the Reuters-21578-LOC-2327-2H data set, again treating each one of the data sets per level as an independent data set.

4.3.6.1 Reuters-21578-LOC-2327-1L

From the results for the Reuters-21578-LOC-2327-1L data set, shown in Table 4.16, it is easy to notice that the best values for both accuracy and AUC were obtained when Chi-squared was used as the feature selection strategy. When the text was preprocessed with Chi-squared feature selection, the highest accuracy values were 82.25% (SMO) and 73.79% (C4.5), whereas the best AUC values were 0.93 (SMO) and 0.87 (Naive Bayes). When CFS feature selection was used, the best accuracy values were 74.82% (SMO) and 65.79% (C4.5), and the best AUC values were 0.89 (SMO) and 0.81 (Naive Bayes). The results show that all the classification methods produced a better performance when the feature selection method used was Chi-squared. Interestingly, for both the Chi-squared and CFS feature selection the best accuracy values were achieved using C4.5 and SMO, whereas the best AUC values were achieved using SMO and Naive Bayes. The classification method that produced the best performance overall was SMO.

4.3.6.2 Reuters-21578-LOC-2327-2L

Table 4.17 presents the results for the Reuters-21578-LOC-2327-2L data set which is very unbalanced and has a very large number of classes (92). All the accuracy and AUC results obtained were higher when Chi-squared feature selection was used, with the exception of that obtained using TFPC-C70 for accuracy and the ones obtained using TFPC-C50, TFPC-C60 and TFPC-C70 for AUC. For both feature selection strategies, the classification methods that resulted in the best performance were C4.5 and SMO. When Chi-squared feature selection was used, the best accuracy results were 73.74% (C4.5) and 74.73% (SMO), and the best AUC results were 0.85 (C4.5) and 0.88 (SMO). Using CFS feature selection the best accuracy results were 69.8 (C4.5) and 70.56 (SMO), whereas the best AUC results were 0.83 (C4.5) and 0.87 (SMO). From the results it can be noticed that most of the classification methods performed better when coupled with Chi-squared feature selection. The best identified classification methods with respect to the Reuters-21578-LOC-2327-2L data set were C4.5 and SMO.

Table 4.16: Classification results for the Reuters-21578-LOC-2327-1L data set with Chi-squared and CFS.

	Chi-squared					CFS				
Method	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
TFPC-C50	56.41	0.12	0.19	0.15	0.95	52.73	0.11	0.17	0.13	0.94
TFPC-C60	56.05	0.12	0.19	0.15	0.95	52.34	0.10	0.16	0.13	0.94
TFPC-C70	55.13	0.12	0.18	0.14	0.95	51.00	0.10	0.14	0.12	0.94
TFPC-C80	53.90	0.11	0.19	0.13	0.94	49.59	0.10	0.14	0.11	0.94
C4.5	73.79	0.83	0.72	0.74	0.92	65.79	0.79	0.64	0.66	0.89
SMO	82.25	0.93	0.81	0.82	0.94	74.82	0.89	0.74	0.75	0.91
LibSVM	72.80	0.78	0.70	0.73	0.84	50.88	0.58	0.49	0.51	0.66
NB	51.01	0.87	0.73	0.51	0.96	43.10	0.81	0.64	0.43	0.95
KNN	62.48	0.74	0.68	0.63	0.86	60.89	0.72	0.61	0.61	0.85
RIPPER	71.85	0.84	0.70	0.72	0.88	64.25	0.78	0.63	0.64	0.81

Table 4.17: Classification results for the Reuters-21578-LOC-2327-2L data set with Chi-squared and CFS.

	Chi-squared					CFS				
Method	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
TFPC-C50	59.72	0.04	0.07	0.07	0.99	59.36	0.03	0.06	0.07	0.99
TFPC-C60	58.91	0.03	0.07	0.07	0.99	58.33	0.03	0.06	0.06	0.99
TFPC-C70	57.37	0.03	0.07	0.07	0.99	57.58	0.03	0.06	0.06	0.99
TFPC-C80	54.34	0.03	0.07	0.06	0.99	50.00	0.02	0.05	0.05	0.99
C4.5	73.74	0.85	0.70	0.74	0.93	69.88	0.83	0.68	0.70	0.92
SMO	74.73	0.88	0.70	0.75	0.89	70.56	0.87	0.67	0.71	0.87
LibSVM	61.93	0.69	0.54	0.62	0.76	43.49	0.52	0.41	0.44	0.60
NB	34.08	0.84	0.55	0.34	0.98	31.20	0.78	0.51	0.31	0.97
KNN	58.23	0.78	0.67	0.58	0.96	57.16	0.68	0.56	0.57	0.86
RIPPER	71.77	0.83	0.66	0.72	0.88	65.62	0.79	0.63	0.66	0.84

4.3.7 Reuters-21578-COM-2327-2H

Tables 4.18 and 4.19 present the results of the two levels of the Reuters-21578-COM-2327-2H data set, again treating each one of the data sets per level as an independent data set.

4.3.7.1 Reuters-21578-COM-2327-1L

The results for the Reuters-21578-COM-2327-1L data set, which is one of the less unbalanced data sets used in this research, are shown in Table 4.18. From the table it is very noticeable that the results were significantly better when Chi-squared feature selection was used. The top two best accuracy results with respect to Chi-squared feature selection were very good: 95.79% (SMO) and 91.41% (RIPPER). The top two best accuracy results obtained using CFS feature selection were 88.48% (SMO) and 84.40% (C4.5). The best AUC result when Chi-squared feature selection was used was 0.99 and it was obtained using SMO and Naive Bayes, whereas when using CFS feature selection 0.95 (SMO) and 0.94 (Naive Bayes) were obtained. Overall Chi-squared feature selection outperformed CFS feature selection. The classification method that produced the best performance in both cases was SMO.

4.3.7.2 Reuters-21578-COM-2327-2L

Table 4.19 shows the results obtained for the Reuters-21578-COM-2327-2L data set, which is very unbalanced and has many classes (52). Similarly to the other data sets that are subsets of the Reuters-21578 data sets, for all the classification methods used the best results were obtained when Chi-squared feature selection was adopted; the best recorded accuracy results were 84.40% (SMO) and 78.99% (C4.5), and the best AUC results were 0.97 (SMO) and 0.94 (Naive Bayes). Using CFS feature selection the best accuracy values were 75.72% (SMO) and 64.68% (C4.5), and the best AUC results were 0.93 (SMO) and 0.87 (RIPPER). Once again SMO was the classification method that produced the best performance in both cases. Similarly to the results produced for the Reuters-21578-LOC-2327-1L data set, the results for the Reuters-21578-COM-2327-2L data set show that all the classification methods produced a better performance when the feature selection strategy used was Chi-squared. As in the case of the other Reuters data sets, the classification method that achieved the best overall results was SMO. The second best methods were C4.5 and RIPPER.

Table 4.18: Classification results for the Reuters-21578-COM-2327-1L data set with Chi-squared and CFS.

	Chi-squared					CFS				
Method	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
TFPC-C50	86.71	0.76	0.82	0.76	0.96	70.94	0.55	0.64	0.55	0.92
TFPC-C60	86.04	0.75	0.82	0.75	0.96	70.12	0.54	0.64	0.54	0.92
TFPC-C70	85.23	0.74	0.82	0.74	0.96	69.01	0.53	0.64	0.54	0.91
TFPC-C80	83.70	0.71	0.80	0.71	0.96	63.21	0.50	0.65	0.50	0.90
C4.5	88.27	0.94	0.88	0.88	0.96	84.40	0.93	0.84	0.84	0.95
SMO	95.79	0.98	0.96	0.96	0.99	88.48	0.95	0.89	0.89	0.96
LibSVM	45.21	0.61	0.49	0.45	0.78	79.24	0.85	0.81	0.79	0.92
NB	89.08	0.98	0.90	0.89	0.97	79.16	0.94	0.82	0.79	0.95
KNN	84.66	0.90	0.86	0.85	0.95	74.13	0.82	0.75	0.74	0.92
RIPPER	91.41	0.96	0.92	0.91	0.97	82.77	0.92	0.85	0.83	0.93

Table 4.19: Classification results for the Reuters-21578-COM-2327-2L data set with Chi-squared and CFS.

	Chi-squared					CFS				
Method	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
TFPC-C50	65.89	0.11	0.16	0.17	0.99	59.86	0.07	0.10	0.11	0.99
TFPC-C60	65.03	0.10	0.15	0.16	0.99	59.72	0.07	0.10	0.11	0.99
TFPC-C70	63.89	0.10	0.15	0.15	0.99	58.80	0.07	0.09	0.10	0.99
TFPC-C80	61.10	0.09	0.14	0.14	0.99	57.41	0.06	0.10	0.10	0.99
C4.5	78.99	0.91	0.77	0.79	0.97	64.68	0.84	0.63	0.65	0.96
SMO	84.40	0.97	0.83	0.84	0.98	75.72	0.93	0.75	0.76	0.96
LibSVM	78.43	0.87	0.72	0.78	0.95	47.40	0.67	0.58	0.47	0.86
NB	49.46	0.94	0.69	0.50	0.99	41.86	0.85	0.59	0.42	0.98
KNN	65.79	0.82	0.77	0.66	0.98	51.48	0.74	0.59	0.52	0.95
RIPPER	78.51	0.92	0.76	0.79	0.96	64.50	0.87	0.67	0.65	0.93

4.3.8 Evaluation Summary

This subsection presents a summary of the evaluation results presented above. The best classification techniques and results with respect to both feature selection methods considered are presented in Table 4.20, along with a description and a comparison of the results. In Table 4.20 the first column presents a listing of all the data sets used. The remaining twelve columns are divided into two groups of six columns, the first group presents the results when Chi-squared feature selection was used and the second group when CFS feature selection was used. For each group the first column lists the names of the best algorithms and the remaining five columns presents the evaluation measures considered (accuracy, AUC, precision, sensitivity/recall and specificity).

Interesting trends can be detected regarding the best classification methods and the data sets to which they were applied. When Chi-squared feature selection was used, for almost all the SAVSNET data sets SMO produced the best results. This was also true in the case of the OHSUMED-CA and Reuters-21578 data sets. For almost all the OHSUMED-AD data sets C4.5 produced the best results. RIPPER obtained the best results only when applied to the SAVSNET-917-2L data set.

On the other hand, when CFS feature selection was used, SMO again produced the best results when it was applied to the OHSUMED-CA and Reuters-21578 data sets. In the case of the SAVSNET data sets, the best algorithms varied between C4.5, RIPPER and SMO. For the OHSUMED-AD data sets, RIPPER tended to produce the best performance.

The criteria for selecting the best classification methods was according to the best accuracy and AUC values, giving more importance to AUC as accuracy can be misleading in the cases where the data sets are very unbalanced. Out of the ten classification methods considered, only three consistently produced a best performance: (i) SMO, (ii) C4.5 and (iii) RIPPER. Although many data sets contained different forms of text and presented different class distributions, the classification method that produced the best performance overall was SMO.

In terms of the feature selection techniques used, in all cases except the SAVSNET-917-2L and SAVSNET-971-3-FT data sets, the best results were obtained when Chi-squared feature selection was adopted. For these data sets the best classification method was the same for both feature selection methods: RIPPER for SAVSNET-917-2L and SMO for SAVSNET-971-3-FT. Comparing the results of the free text only and the free text combined with tabular data variations of the SAVSNET-840-4 and SAVSNET-971-3 data sets, it was interesting that when Chi-squared feature selection was used for both data sets, the inclusion of tabular data in the free text improved the classification results. On the other hand, when CFS feature selection was used, for both data sets the highest results were obtained when only free text was considered.

Table 4.20: Best classification techniques and results.

Data set	Chi-squared						CFS					
	Best algorithm	Best results					Best algorithm	Best results				
		Acc (%)	AUC	Pr	Sn/Re	Sp		Acc (%)	AUC	Pr	Sn/Re	Sp
SAVSNET-840-4-FT	SMO	89.29	0.95	0.90	0.89	0.96	C4.5	88.33	0.94	0.88	0.88	0.95
SAVSNET-840-4-TD+FT	SMO	90.60	0.96	0.91	0.91	0.96	RIPPER	85.33	0.94	0.85	0.86	0.93
SAVSNET-971-3-FT	SMO	74.87	0.80	0.74	0.75	0.79	SMO	76.21	0.79	0.75	0.76	0.78
SAVSNET-971-3-TD+FT	SMO	75.30	0.80	0.75	0.75	0.81	C4.5	68.70	0.75	0.66	0.69	0.70
SAVSNET-917-1L	SMO	70.34	0.75	0.68	0.70	0.75	SMO	67.61	0.72	0.65	0.68	0.71
SAVSNET-917-2L	RIPPER	66.96	0.57	0.60	0.67	0.45	RIPPER	68.38	0.58	0.63	0.68	0.47
SAVSNET-917-3L	C4.5	57.25	0.59	0.56	0.57	0.57	C4.5	56.27	0.57	0.54	0.56	0.55
SAVSNET-917-4L	SMO	47.00	0.64	0.47	0.47	0.75	SMO	45.37	0.59	0.42	0.45	0.69
OHSUMED-CA-3187-1L	SMO	94.84	0.93	0.95	0.95	0.91	SMO	94.30	0.91	0.94	0.94	0.88
OHSUMED-CA-2570-2L	SMO	80.82	0.94	0.81	0.81	0.96	SMO	74.46	0.92	0.74	0.75	0.95
OHSUMED-CA-834-3L	SMO	78.42	0.90	0.79	0.78	0.92	SMO	76.29	0.89	0.75	0.76	0.92
OHSUMED-AD-3393-1L	SMO	82.46	0.85	0.81	0.83	0.77	RIPPER	74.17	0.71	0.71	0.74	0.62
OHSUMED-AD-569-2L	C4.5	76.74	0.87	0.78	0.77	0.93	RIPPER	73.17	0.93	0.82	0.77	0.97
OHSUMED-AD-292-3L	C4.5	89.69	0.91	0.90	0.90	0.93	C4.5	69.07	0.82	0.68	0.69	0.80
Reuters-21578-LOC-2327-1L	SMO	82.25	0.93	0.81	0.82	0.94	SMO	74.82	0.89	0.74	0.75	0.91
Reuters-21578-LOC-2327-2L	SMO	74.73	0.88	0.70	0.75	0.89	SMO	70.56	0.87	0.67	0.71	0.87
Reuters-21578-COM-2327-1L	SMO	95.79	0.98	0.96	0.96	0.99	SMO	88.48	0.95	0.89	0.89	0.96
Reuters-21578-COM-2327-2L	SMO	84.40	0.97	0.83	0.84	0.98	SMO	75.72	0.93	0.75	0.76	0.96

4.4 Text Summarisation

This section presents an evaluation of the summaries generated. Examples of generated summaries for each data set are shown in Table 4.21. The first column records the name of the data set, the second column indicates the ID of the record within the respective data set, the third column contains the original free text (without preprocessing), the fourth column shows the name of the class which was assigned to the document and the fifth column shows the generated summary based on the name of the assigned class. The purpose of showing the free text without preprocessing is to emphasize the challenge that was involved in automating the summarisation process.

Given that in some cases there could be a large number of classes in a data set and that the names of such classes are usually long, during the internal text classification process, the name of the class was coded using a two or three character code (e.g. “Diarrhoea” was coded as “AA”) in order to have a more standard representation of the classes and to simplify their handling. Using standard classification only a single class label is taken into account when generating summaries. Consequently the generated summaries may be considered to not be as complete as might be preferred. The main advantages of the technique presented in this chapter with respect to the summarisation process are:

1. The resulting summary is a concise way to present the main idea of what the text is about.
2. It is not as computationally expensive and complex as a summary generated with multi-label classification.

Recall that the resulting summaries are constructed by prepending or appending a domain-specific sentence to the name of the class. For example, in the summary “Diarrhoea was presented.” the name of the class (“Diarrhoea”) starts the sentence followed by a domain-specific phrase related to the name of the class (“was presented”). As already noted, the quality of the generated summaries depends very much on the quality of the classification process, thus having a good classifier will result in a good classification of documents and consequently a good summary.

Since the text summarisation approaches presented in this thesis rely on text classification methods, linguistics and the meaning and order of the words in the text is not considered. Instead, only the names of the classes to which the documents relate are utilised, resulting in a substantial reduction of the summary generation time.

The generated summaries were evaluated in terms of both the intrinsic and extrinsic evaluation measures typically used for text summarisation, which were presented in Chapter 2. In the case of the former, the summaries complied with all the intrinsic evaluation criteria as they were: grammatically correct, non-redundant, complete,

accurate, structured and coherent and presented referential clarity. In terms of the extrinsic evaluation measures, on the other hand, the summaries were considered as acceptable considering that the results were obtained with text classification, which was the only quantitative evaluation measure used. Regarding the information retrieval and the question answering qualities of the summaries, if this technique was to be integrated into a piece of software for a specific domain, it could prove to be very useful with respect to users who wish to retrieve the content of unseen documents according to the generated text summaries.

Table 4.21: Examples of summaries generated using standard classification techniques.

Data set	ID	Free text	Class	Summary
SAVSNET-840-4	34386	Weight = 4.24 kgs. SAVSNET Vomit and Diarrhoea survey completed Repeat Consultation Injection sub-cut/i-musc 0.50 mls Voren Suspension (50ml) 0.50 x Betamox LA inj 100ml (mls) 10 x Stomorgyl 2 Tablets (tabs) Give 1 tablet(s) twice daily Still D++ Bright alert and responsive app+ etc and No abnormalities detected on exam Faecal sample provided for analysis Start Treatment in meantime and delay vacs for 1 week pending results and response to Treatment Lab Test - Gastro 9 (faeces) Submission/Handling/Interpretation	Diarrhoea	“Diarrhoea was presented.”
SAVSNET-840-4-TD+FT	34386	canine bassethound unknownSex unneutered tricolour notMicrochipped uninsured notDeceased Weight = 4.24 kgs. SAVSNET Vomit and Diarrhoea survey completed Repeat Consultation Injection sub-cut/i-musc 0.50 mls Voren Suspension (50ml) 0.50 x Betamox LA inj 100ml (mls) 10 x Stomorgyl 2 Tablets (tabs) Give 1 tablet(s) twice daily Still D++ Bright alert and responsive app+ etc and No abnormalities detected on exam Faecal sample provided for analysis Start Treatment in meantime and delay vacs for 1 week pending results and response to Treatment Lab Test - Gastro 9 (faeces) Submission/Handling/Interpretation	Diarrhoea	“Diarrhoea was presented.”

Table continues in the next page.

Table 4.21 continued.

Data set	ID	Free text	Class	Summary
SAVSNET-971-3	358	SAVSNET Vomit and Diarrhoea/// [211]survey completed/// [211]Repeat Consultation/// [164]Injection sub-cut/i-musc/// [98]2 mls Voren Suspension (50ml)/// [12]2 x Betamox inj 100ml (mls)/// [10]14 x Metronidazole 200mg (tabs) Give 1 tablet(s) twice daily/// [10]7 x Zantac 75mg (per tab) Give 1 tablet(s) once daily/// [14]owner reports colitis symptoms intermittent over 2-3 weeks On exam No abnormalities detected T=normal. owner stopped metacam 3 days ago. Start Treatment as above for colitis and restart metacam in 3-4 days Adv light diet for 48hrs/// [200]See if owner worried/// [200]	Diarrhoea	“Diarrhoea was presented.”
SAVSNET-971-3-TD+FT	358	canine crossbreed aged unknownSex neutered SAVSNET Vomit and Diarrhoea/// [211]survey completed/// [211]Repeat Consultation/// [164]Injection sub-cut/i-musc/// [98]2 mls Voren Suspension (50ml)/// [12]2 x Betamox inj 100ml (mls)/// [10]14 x Metronidazole 200mg (tabs) Give 1 tablet(s) twice daily/// [10]7 x Zantac 75mg (per tab) Give 1 tablet(s) once daily/// [14]owner reports colitis symptoms intermittent over 2-3 weeks On exam No abnormalities detected T=normal. owner stopped metacam 3 days ago. Start Treatment as above for colitis and restart metacam in 3-4 days Adv light diet for 48hrs/// [200]See if owner worried/// [200]	Diarrhoea	“Diarrhoea was presented.”

Table continues in the next page.

Table 4.21 continued.

Data set	ID	Free text	Class	Summary
SAVSNET-917-1L	122150	Pro-Kolin 15ml 2ml three times daily [[254]5 x Sulphasalazine 1/4tablet twice daily [[254]Inject S/C 0.50 x Dexafort inj [[254]SAVSNET Vomit and Diarrhoea [[211]survey completed [[211]Consultation 2 [[2]dial on and off for 1yr since bout when on hols Waxes and wanes but always a bit loose VC well in self TN No wt loss Prob underlying ?sensitivity + colitis Adv re diet and treat If no imp ?faec sample PS [[200]	Diarrhoea	“Diarrhoea was presented.”
SAVSNET-917-2L	122170	WEIGHT = 2.90 kgs. [[229]SAVSNET Vomit and Diarrhoea [[211]survey completed [[211]Consultation [[1]Inject 0.30 mls CERENIA [[254]chronic V+ x1/wk 1 yr last few days more frequent & fresh blood this morn HR=220 & 1/6 murmur no goitre nad abdo palp temp normal adv re poss causes starve & white fish INB adv BT FHP & T4-if all OK may try steroid_o wouldnt want [[200]extensive investigations at this point O has wormed [[200]	Haemorrhagic	“The condition presented was haemorrhagic.”
SAVSNET-917-3L	123152	Weight 10.10 kgs. [[229]SAVSNET Vomit and Diarrhoea [[211]survey completed [[211]Examination by Vet [[12]inject 0.50 mls Synulox injection [[2]3 x Hills canine ID tin (tins) 1/4 tin 2-3 times daily [[7]5 x Noroclav 250mg tablets half tablet twice daily [[3]o reports started with D+ today contained blood_ V+ once today_ frothy material_ off colour today. temp 101.8F adv symptomatic tx but rv if no imp in 24-48hrs. o wants to book in for dental_ adv leacve until V+ and D+ resolve [[200]inject ml Voren injection [[2]	First Time	“The condition was presented for the first time.”

Table continues in the next page.

Table 4.21 continued.

Data set	ID	Free text	Class	Summary
SAVSNET-917-4L	123199	Weight 34.00 kgs. [[229]SAVSNET Vomit and Diarrhoea [[211]survey completed [[211]Consultation (Standard) [[50]Pharmacy Diarsanyl+ 60ml (Lge Dog) Give 1 divs2. times daily for 7days Vetaf [[254]metronidazole 200mg prescribed [[100]Pharmacy 30 x Metronidazole 200mg (Give 3 tab/s 2 times/day for 5 days Vetaf [[254]Discount: Pet Aid [[213]d+ for the past 4 days. bright in herself and still keen to eat. abdo relaxed-temp 40 although really scared in here. advise starve then abov tx. o says she doesn't tolerate chicken well so he will give her boiled white fish [[138]and rice. [[138]Donation Received (PDSA)- 2.50 [[196]Pediatric Doppler echocardiography 1987: major advances in technology. The major advances in the capabilities of pediatric cardiologists to evaluate the heart by ultrasound that have occurred in the last 5 years have been reviewed. In addition to the new Doppler methods, the evolution of higher resolution echo techniques have provided a comprehensive means of evaluating the heart noninvasively. This information has relegated catheterization to a more therapeutic arena, leaving ultrasound as the major diagnostic technique for evaluation of congenital heart disease, both before and after birth.	Between Two And Four Days	"The condition was presented between two and four days."
OHSUMED-CA-3187-1L	87140544		Congenital Heart Defects	"The document is about Congenital Heart Defects."

Table continues in the next page.

Table 4.21 continued.

Data set	ID	Free text	Class	Summary
OHSUMED-CA-2570-2L	88066344	Anterior inferior cerebellar artery aneurysm, carotid bifurcation aneurysm, and dural arteriovenous malformation of the tentorium in the same patient. An exceptional combination of intracranial vascular malformations is reported: distal anterior inferior cerebellar artery (AICA) aneurysm, carotid bifurcation aneurysm, and dural arteriovenous malformation (DAVM) of the tentorium. The AICA aneurysm was the source of recurrent subarachnoid and cerebellar hemorrhage, revealed only after repeated vertebral angiography. After external drainage of associated hydrocephalus, both aneurysms were successfully clipped and the dural malformation was subtotally embolized. The literature concerning AICA aneurysms, DAVMs, and combined intracranial vascular malformations is reviewed and discussed.	Arteriovenous Malformations	“The document is about Arteriovenous Malformations.”
OHSUMED-CA-834-3L	90263077	Balloon embolization of iatrogenic aortocoronary arteriovenous fistula. A detachable latex balloon was used to occlude an iatrogenic aortocoronary arterio-venous fistula. The aim of providing retrograde myocardial perfusion was not achieved to any significant degree because of rapid recruitment of collateral venous routes to the coronary sinus. This may have implications for the effectiveness of deliberate grafting of the coronary venous system with proximal venous ligation, as has been recommended when the coronary arterial system is small and diffusely diseased.	Arteriovenous Fistula	“The document is about Arteriovenous Fistula.”

Table continues in the next page.

Table 4.21 continued.

Data set	ID	Free text	Class	Summary
OHSUMED-AD-3393-1L	89076819	Epithelial cutaneous lesions induced in Dunkin-Hartley albino guinea-pigs by means of 7,12-dimethyl-benzanthracene. We carried out a clinicopathological study of epithelial cutaneous lesions induced in Dunkin-Hartley albino guinea-pigs by means of the topical application of 7,12-dimethyl-benzanthracene. By the end of the study we had observed 4451 lesions. Most frequently observed were epithelial hyperplasias and dysplasia (2544 lesions). Neoplasias consisted of 244 carcinomas in situ, and 88 squamous carcinomas. Basal-cell epitheliomas were few. Epidermal cysts were a frequent finding. Protracted enteric cryptosporidial infection in selective immunoglobulin A and saccharomyces opsonin deficiencies [see comments] Chronic cryptosporidial infection in man usually occurs in those who are immunocompromised. We report a patient with a one year history of bowel symptoms resulting from persistent cryptosporidial infection of the colon. Investigations showed underlying selective IgA and saccharomyces opsonin deficiencies but no evidence of cell mediated immune dysfunction. Both selective immunoglobulin A and opsonin deficiencies are relatively common in the general population and may be a cause of susceptibility to persistent cryptosporidial infection.	Animal Disease Models	“The document is about Animal Disease Models.”
OHSUMED-AD-569-2L	90337437		Animal Protozoan Infections	“The document is about Animal Protozoan Infections.”

Table continues in the next page.

Table 4.21 continued.

Data set	ID	Free text	Class	Summary
OHSUMED-AD-292-3L	91298715	Effect of clindamycin on pneumonia from reactivation of Toxoplasma gondii infection in mice. Clindamycin was used to treat the reactivation of a chronic Toxoplasma gondii infection in mice. Clindamycin reduced mortality by 44% when used prophylactically (P less than 0.001) but appeared to be less effective when used to treat clinically apparent reactivation. Further studies should be conducted to establish the efficacy of clindamycin for the treatment of toxoplasmosis in humans.	Animal Toxoplasmosis	"The document is about Animal Toxoplasmosis."
Reuters-21578-LOC-2327-1L	1600	CCC ACCEPTS EXPORT BID FOR WHEAT FLOUR TO IRAQ WASHINGTON, April 8 - The Commodity Credit Corporation has accepted a bid for an export bonus to cover a sale of 12,500 tonnes of wheat flour to Iraq, the U.S. Agriculture Department said. The bonus awarded was 113.0 dlrs per tonne and will be paid to Peavey Company in the form of commodities from CCC stocks. The wheat flour is for delivery May 15-June 15, 1987, the department said. An additional 162,500 tonnes of wheat flour are still available to Iraq under the Export Enhancement Program initiative announced January 7, 1987, USDA said. Reuters	AmericaAsia	"This document has news related to America and Asia."

Table continues in the next page.

Table 4.21 continued.

Data set	ID	Free text	Class	Summary
Reuters-21578-LOC-2327-2L	1129	<p>PEGASUS GOLD ;PGULF; STARTS MILLING IN MONTANA JEFFERSON CITY, Mont., March 27 - Pegasus Gold Inc said milling operations have started at its Montana Tunnels open-pit gold, silver, zinc and lead mine near Helena. The start-up is three months ahead of schedule and six mln dlrs under budget, the company said. Original capital cost of the mine was 57.5 mln dlrs, but came in at 51.5 mln dlrs, the company said. After a start-up period, the mill is expected to produce 106,000 ounces of gold, 1,700,000 ounces of silver, 26,000 tons of zinc and 5,700 tons of lead on an annual basis from 4,300,000 tons of ore, the company said, Reuters</p> <p>PERU SAYS NEW GOLD DEPOSITS WORTH 1.3 BILLION DLRS LIMA, March 29 - President Alan Garcia said Peru has found gold deposits worth an estimated 1.3 billion dlrs in a jungle region near the Ecuadorean border about 1,000 km north of here. He told reporters the deposits, located at four sites near the town of San Ignacio, contained the equivalent of 100 tonnes of gold. Garcia said the government would soon install a two mln dlr treatment plant at Tomaque. It will extract enough ore to provide an estimated 25 mln dlr profit by the end of this year, he added. Garcia said the other gold-bearing deposits are located at Tamborapa, Pachapidiana and a zone between the Cenepa and Santiago rivers. REUTERS</p>	USA	“This document has news related to USA.”
Reuters-21578-COM-2327-1L	1154	<p>PERU SAYS NEW GOLD DEPOSITS WORTH 1.3 BILLION DLRS LIMA, March 29 - President Alan Garcia said Peru has found gold deposits worth an estimated 1.3 billion dlrs in a jungle region near the Ecuadorean border about 1,000 km north of here. He told reporters the deposits, located at four sites near the town of San Ignacio, contained the equivalent of 100 tonnes of gold. Garcia said the government would soon install a two mln dlr treatment plant at Tomaque. It will extract enough ore to provide an estimated 25 mln dlr profit by the end of this year, he added. Garcia said the other gold-bearing deposits are located at Tamborapa, Pachapidiana and a zone between the Cenepa and Santiago rivers. REUTERS</p>	Metal	“This document has news related to metal.”

Table continues in the next page.

Table 4.21 continued.

Data set	ID	Free text	Class	Summary
Reuters-21578-COM-2327-2L	1179	<p>TRADE SEES U.S. CORN EXPORTS UP, WHEAT/BEANS OFF CHICAGO, March 30 - Grain traders and analysts expect lower wheat and soybean exports and higher corn exports than a year ago in the USDA's export inspection report today. Corn export guesses ranged from 27.0 mln to 32.0 mln bushels, compared with the 27.6 mln inspected last week and 20.5 mln a year ago. Soybean export guesses ranged from 14.0 mln to 16.0 mln, up from the 13.4 mln inspected last week but below the 25.5 mln reported a year ago. Wheat estimates ranged from 11.0 mln to 14.0 mln bushels, compared with 12.0 mln reported last week and 18.3 mln a year ago. Reuters</p>	Grain	<p>"This document has news related to grain."</p>

4.5 Discussion

This section presents a discussion regarding the obtained results and how they performed with respect to the objectives presented at the beginning of this chapter:

1. From the results obtained in this chapter, a benchmark has been established for use with respect to the techniques described in Chapters 5, 6 and 7.
2. The operation of a number of classifiers using data sets containing different forms of text was compared and the best ones (SMO, C4.5 and RIPPER) were identified. Note that the SAVSNET, OHSUMED and Reuters data sets contain different forms of free text: (i) questionnaire free text, (ii) academic text and (iii) text form news articles. In most cases, the classification method that consistently produced the best performance was SMO. The other two best standard classification methods were C4.5 and RIPPER. In general the results obtained with all the classification methods were relatively good, considering the unbalanced nature of some of the data sets.
3. For the SAVSNET-840-4 and SAVSNET-971-3 questionnaire data sets, which included both tabular and free text, the experiments were carried out using: (i) the free text only and (ii) both the free text and tabular data combined. In most cases there was no indication as to whether the inclusion of tabular data improved the classification or not. The exception to this was in the case of the SAVSNET-840-4-TD+FT data, when Chi-squared feature selection was used, where combining free text and tabular data worked well. In the case of SAVSNET-971-3-FT, with CFS feature selection, the best results obtained with the free text only data were substantially higher than the ones obtained by combining free text and tabular data. In terms of the best results, the feature selection method adopted did not seem to have a significant role in the improvement of the results. Note that these results were obtained using only the SAVSNET data set, therefore more experiments might be needed to establish more generally that: (i) the inclusion of tabular data in free text improves classification results when Chi-squared feature selection is used and (ii) that considering only free text improves the classification results when CFS feature selection is used (this is discussed further in Chapter 8). With respect to the remainder of the work described in this thesis it was thus decided to focus only on the free text part of questionnaire data even if additional tabular data was available.
4. In most cases the best results were obtained when Chi-squared was used as the feature selection strategy. In many cases the results obtained using CFS feature selection were close to those obtained using Chi-squared feature selection. How-

ever, in the context of the alternative techniques described in this thesis, where appropriate, only Chi-squared feature selection was adopted.

Overall, using standard classification techniques for the purpose of summary generation seemed to produce acceptable outcomes and thus it is argued that a suitable benchmark had been established.

4.6 Summary

This chapter considered the appropriateness of using standard classification techniques to generate summaries. Seven standard classification techniques, representative of the main predictive models found in the data mining literature, were applied to the data sets introduced in Chapter 3, namely: (i) Naive Bayes, (ii) C4.5, (iii) TFPC, (iv) RIPPER, (v) KNN, (vi) SMO and (vii) LibSVM. Four variants of TFPC were used, using different confidence threshold values. In addition to using different standard classification techniques, two different feature selection methods were also considered, namely: (i) Chi-squared and (ii) CFS. The evaluation metrics used were: (i) overall accuracy expressed as a percentage, (ii) Area Under the ROC Curve (AUC), (iii) precision, (iv) sensitivity/recall and (v) specificity. Accuracy and AUC were the main evaluation methods considered; while precision, sensitivity/recall and specificity were used to provide a broader insight of the classification results. Stratified Ten-fold Cross Validation (TCV) was used with respect to all the reported experiments.

The quality of the summaries in each case depended on how well the classifier performed, and also according to which feature selection method was adopted. Except with respect to the hierarchical summarisation technique proposed in Chapter 7, the summary generation mechanisms presented in this thesis are all based on single class labels. This offers the advantages that: (i) the resulting summaries are a concise way of presenting the main idea of what a given piece of text is about and (ii) it is not as computationally expensive and complex as a summary generated with (say) multi-label classification. In addition some examples of the obtained text summaries were presented and discussed.

In conclusion, the best classification techniques were SMO, C4.5 and RIPPER. The best feature selection technique was Chi-squared feature selection. The use of tabular data, where available, was found not to improve the quality of the classification and consequently the summaries generated. A benchmark was established for use with respect to further experimentation.

Chapter 5

Classifier Generation Using Secondary Data (CGUSD) for Text Summarisation

5.1 Introduction

The previous chapter considered the appropriateness of using standard classification techniques to generate summaries and established a benchmark for the techniques presented in this chapter and in Chapters 6 and 7. This chapter presents a technique that considers the case when it is desired to provide a summary from a questionnaire corpus (or any other type of text corpus), regarding the nature of the free text element of such questionnaires, where the available data is not considered suitable for training purposes (or possibly because no data is available at all). It is common that primary data sets can be limited in terms of number of records and suitability for various reasons such as the scarcity of labelled documents on a certain subject (class) or the confidentiality of the documents related to sensitive data (personal information, national security, etc.) for example. The approach advocated in this chapter is to build the desired text summarisation classifier using appropriate secondary data and then applying the resulting classifier, for the purpose of text summarisation, to the primary data (free text that is to be summarised). It is desired to search, extract and generate appropriate secondary data in order to build the desired text summarisation classifier. For this to operate successfully the secondary data must feature the same topics (class labels) as the primary data. The approach presented in this chapter is referred to as the CGUSD (Classifier Generation Using Secondary Data) approach. The research described in this chapter had four objectives:

1. To determine whether the desired text summarisation classifiers could be generated using secondary data, and whether the resulting classifiers could then be successfully applied to primary data so as to produce adequately the desired summarisations (assuming that appropriate secondary data is available).

2. To ascertain the applicability and effectiveness of the CGUSD approach when applied to free text from different sources than questionnaires.
3. To assess the quality of the summarisation classifiers generated considering that the secondary data sets can be constructed (in most cases) in a balanced manner such that there is an even distribution of records with respect to each class.
4. To compare the operation of the three classifiers that had the best performance in Chapter 4 (SMO, C4.5 and RIPPER) with respect to the CGUSD approach.

Note that with respect to the third objective an issue frequently encountered in data mining is the issue of unbalanced data sets [12]. It is expected that a sufficient number of examples will not be available in all cases, for some class labels there may be a large number of records (for example classes that represent common occurrences), for others there may be very few (for example classes that represent rare occurrences). Therefore input data can generally be expected to be *unbalanced*. In the context of data mining and machine learning the terms “imbalanced” and “unbalanced” are used interchangeably. In this thesis the term “unbalanced” is used. Unbalanced data sets typically cause text classification learning algorithms to become biased towards the classes that have the majority of examples; consequently, misleading classifications with respect to unseen documents may be produced.

The remainder of this chapter is arranged as follows. The proposed CGUSD methodology is described in Section 5.2. A description of the secondary data sets, and how they relate to the primary data sets used, is presented in Section 5.3. The experiments carried out to evaluate CGUSD are reported on in Section 5.4. A discussion of how the proposed technique performed is presented in Section 5.5. Finally, a summary of the chapter is presented in Section 5.6.

5.2 The Classifier Generation Using Secondary Data (CGUSD) Methodology

The CGUSD methodology is illustrated in Figure 5.1. The methodology encompasses five stages: (i) secondary data generation, (ii) preprocessing and representation (using a vector space model) of both the primary and secondary data sets, (iii) feature selection with respect to the secondary data, (iv) classification and (v) summarisation. The operation of the second and third stages was described extensively in Chapter 3. However, in the case of CGUSD both the primary and secondary data sets must be pre-processed in the same manner and represented separately in terms of a feature vector representation during the second and third stages. Text summarisation is carried out in the same way as in Chapter 4, and is thus not commented on further in this chapter. The first and the fourth stages (shown in Figure 5.1 in the areas labelled “(1) Secondary

data set generation” and “(4) Classification” respectively) are the most significant with respect to the proposed CGUSD approach because they have a major influence on the effectiveness of the resulting classifier and the consequent text summarisation. These two stages are therefore described in detail in Subsections 5.2.1 and 5.2.2.

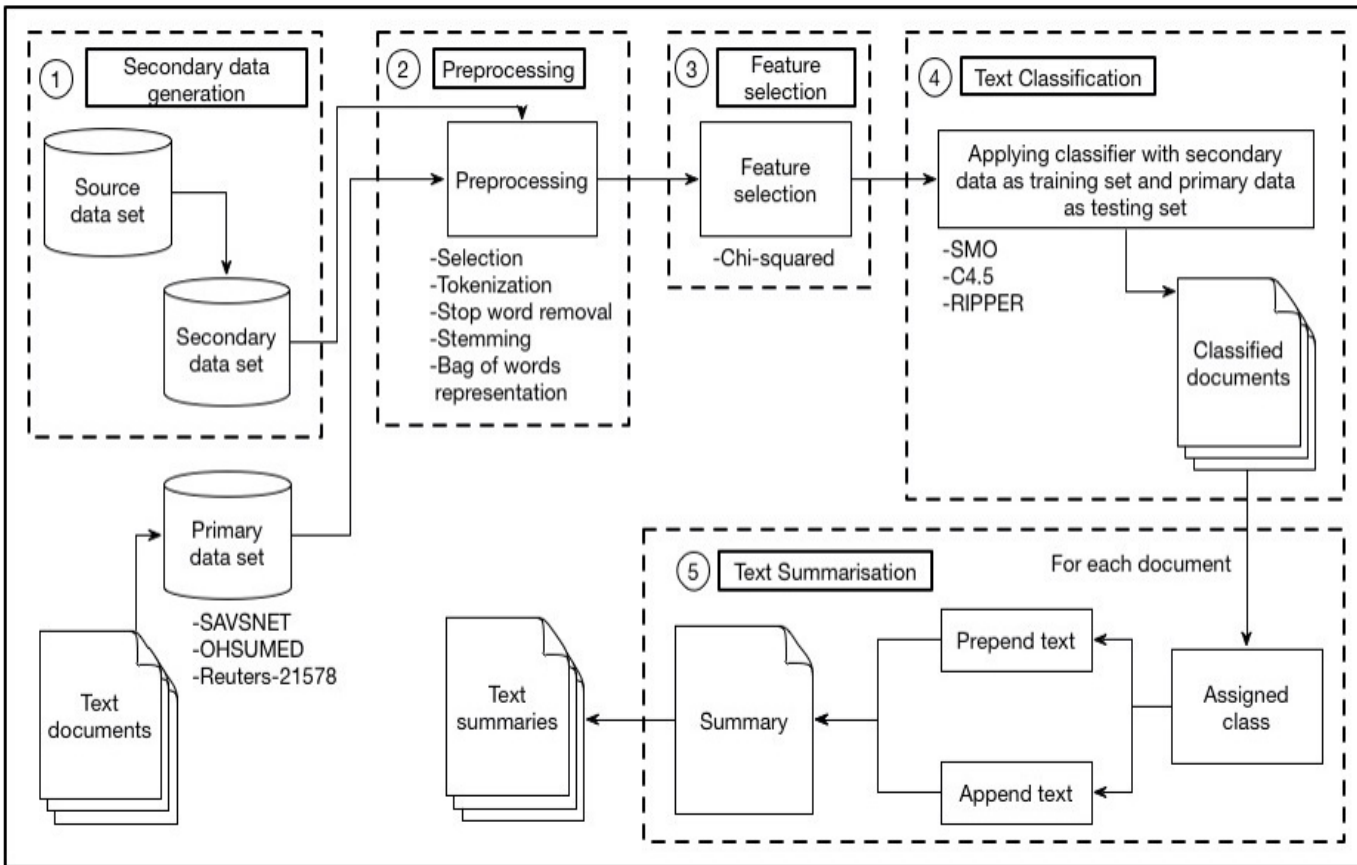


Figure 5.1: Classifier Generation Using Secondary Data (CGUSD) for Text Summarisation.

5.2.1 Secondary data set generation

In order to generate adequate secondary data it is necessary that the source data is compatible with the primary data in terms of the nature of the data and the class labels used, which in most cases it is not a straightforward process. The author identifies two types of source data in the form of publicly available repositories that can be used to extract secondary data: (i) curated, cleaned and indexed data set repositories, and (ii) repositories where data is indexed and catalogued but presented in raw format. In the context of data mining research, the former source of data is commonly used in the literature to compare the operation of data mining approaches with different data sets. The latter source of data is also used in data mining research, but preprocessing needs to be done in order to transform the data sets into the desired format depending on the application in which they will be used. The SAVSNET collection fell into the second category, because it was indexed and catalogued, but the free text was in raw format. The OHSUMED and Reuters-21578 collections fell into the first category because although they were curated, cleaned and indexed, they were presented in XML format. For all the data collections used the free text had to be transformed into the appropriate format for SARSET.

Once it has been acknowledged that the primary data is not adequate to generate text summarisation classifiers, it is necessary to find a reliable source of data closely related to the area or topic of the primary data. The class labels from the primary data must be present in the source data either: (i) implicitly (as a result of combining other class labels) or (ii) explicitly (having already the required class labels). The secondary data source should also be such that a sufficient quantity of secondary data can be obtained, there is little point in obtaining a secondary data set if this data set is inferior to the original primary data set. The proposed procedure to extract the optimal and balanced number of secondary documents to be used to generate text summarisation classifiers is described below.

From Figure 5.1 it can be observed that the input to the CGUSD methodology comprised both a primary data set P and secondary data set S . The latter is generated from an alternative data source related to the primary data. The secondary data generation process, as shown in the area labelled “(1) Secondary data generation” in Figure 5.1, involves interacting with the alternative source from which the secondary data will be extracted; thus, with respect to the work described in this thesis, some kind of document collection. Recall that it is desired to build a multi-class classifier given a set of n class labels $C = \{C_1, C_2, \dots, C_n\}$. Hence we want a sufficient number of examples from the secondary data so that each class is appropriately represented. The secondary data is represented as $S = \{S_1, S_2, \dots, S_n\}$ where each element S_i represents the set of documents associated with class C_i . It is also desirable that the number of documents in each set S_i is the same so that we have a balanced data set so as to ensure

that the classifier generation process does not favour a majority class.

$$\forall S_i \in S, |S_i| = k$$

where k is a constant.

Experience shows that a better classifier is generated if the number of documents in S is such that all potential cases are “covered”, thus k needs to be of a reasonable size. However, the author has also discovered that in practice, depending on the nature of the application domain and the nature of the secondary data source, it is not always possible to generate a secondary data set that is both balanced and of a reasonable size.

In the author’s work presented in [34] the value of k was defined by the user (domain expert) and was simply interpreted as a maximum, if k documents could not be retrieved with respect to a particular class a lower number was accepted. Since the results obtained in [34] were not as promising as expected an alternative secondary data generation process was adopted. The revised process is presented in Algorithm 2.

Data: $P = \{P_1, P_2, \dots, P_n\}$, $S = \{S_1, S_2, \dots, S_n\}$, k
Result: a balanced secondary data set D larger than the primary data set P

```

1  $D = \emptyset$ 
2 for all  $S_i$  in  $S$  from  $i = 1$  to  $i = n - 1$  do
3   if  $|S_i| \leq k$  then
4      $S' = \{S'_1, S'_2, \dots, S'_n\}$  ( $S$  ordered according to size)
5     for all  $S'_i$  in  $S'$  from  $i = 1$  to  $i = n - 1$  do
6        $S'' = \{S''_1, S''_2, \dots, S''_n\}$  ( $S'$  pruned according to  $|S'_i|$ )
7       if  $|S''_i| \leq |P|$  then
8         appropriately sized secondary data set can not be extracted
9         exit;
10      else
11         $D = |S''|$ 
12      end
13    end
14  else
15     $D = D \cup S_i$ 
16  end
17 end
18 return  $D$ 

```

Algorithm 2: Secondary data set generation.

The input to the algorithm is a set P , a set S and a value for k . The expected result is a balanced secondary data set D larger than the primary data set P . The candidate secondary data set $S = \{S_1, S_2, \dots, S_n\}$ is generated using a user defined maximum number of documents per class k . The value for k is selected so that it exceeds the expected maximum number of documents that can be retrieved for any given class.

This requires a certain amount of domain knowledge and expertise. The primary data set $P = \{P_1, P_2, \dots, P_n\}$ contains subsets related to each of the classes defined for P , thus one subset P_i per class C_i . Firstly, the balanced secondary data set D is defined as an empty set (line 1). The algorithm then loops through S (lines 2 - 17). On each iteration:

1. Each subset, $|S_i|$ in S is compared with k :

$$\forall S_i \in S, |S_i| \leq k$$

If $|S_i|$ is less than k , then the set S is arranged in ascending order of size of each of the component subsets to give: $S' = \{S'_1, S'_2, \dots, S'_n\}$ (line 4), otherwise S_i is added to D so far (line 15).

2. The algorithm loops through S' and for each set S'_i in S' a set S'' is generated by pruning S' so that each element in S' has the size $|S'_i|$ (line 6).
3. Next, the size of S'' is revised by iteratively comparing it to the size of P . If the size of S'' is not greater than the size of P it is acknowledged that an appropriate secondary data set D cannot be identified and therefore CGUSD is not applicable (exit the process). Otherwise, the secondary data set D is generated from $|S''|$.
4. If $|S_i|$ is greater than k , then D takes the resulting set of the union of the empty set D and S_i (line 15).
5. Finally, a balanced secondary data set D larger than the primary data set P is returned.

5.2.2 Classification

The classification stage is shown in the area labelled “(4) Text Classification” in Figure 5.1. At this stage the secondary data has been generated and both the primary and the secondary data sets have been processed so that each is represented using a set of feature vectors. In the classification stage the classifier generation takes place using the secondary data as the training set. For evaluation purposes (see Section 5.4) three classifier generation mechanisms were considered, those that produced the best performance as established by the experiments reported in Chapter 4, namely: (i) SMO, (ii) C4.5 and (iii) RIPPER. In the author’s work presented in [34] only the TFPC algorithm was considered and the results obtained were not as good as expected mainly, as the author conjectured, because: (i) the manner in which the secondary data was extracted and preprocessed, and (ii) TFPC might not have been the most appropriate classification technique.

5.3 Secondary data sets used

The secondary data sets used with regard to the primary data sets used for evaluation purposes in this thesis are described in this section. The primary data sets used were:

- SAVSNET-840-4-FT
- SAVSNET-971-3-FT
- SAVSNET-917-1L
- OHSUMED-CA-3187-1L
- OHSUMED-CA-2570-2L
- OHSUMED-CA-834-3L
- OHSUMED-AD-3393-1L
- OHSUMED-AD-569-2L
- OHSUMED-AD-292-3L
- Reuters-21578-LOC-2327-1L
- Reuters-21578-LOC-2327-2L
- Reuters-21578-COM-2327-1L

Six of the original data sets (SAVSNET-840-4-TD+FT, SAVSNET-971-3-TD+FT, SAVSNET-917-2L, SAVSNET-917-3L, SAVSNET-917-4L, Reuters-21578-COM-2327-2L) are not considered in this chapter because, either: (i) the class labels in the primary data sets were too specific to be related to any readily available secondary data set, or (ii) suitable but not sufficient data was available, or (iii) data sets that included a tabular data element were excluded (because work presented in Chapter 4 had established that the inclusion of the tabular data did not provide any benefit). Thus the number of data sets considered in this chapter was reduced from 18 to 12.

Recall also that the evaluation data sets can be grouped under the following headings: (i) SAVSNET, (ii) OHSUMED and (iii) Reuters-21578. Each of these groupings can be associated with particular types of secondary data. SAVSNET and OHSUMED are data sets related to both the veterinary and human medical areas, so it was deemed appropriate to use the MEDLINE (Medical Literature Analysis and Retrieval System Online) database as the source of the secondary data in both cases. The Reuters-21578 data sets are related to news stories and hence the RCV1 (Reuters Corpus Volume 1) was chosen as the source of the secondary data. In the following two subsections MEDLINE and RCV1 are described and discussed in relation to the primary data sets.

5.3.1 MEDLINE (Medical Literature Analysis and Retrieval System Online)

MEDLINE (Medical Literature Analysis and Retrieval System Online) is a life science and biomedical bibliographic database maintained by the United States National Library of Medicine (NLM)¹. MEDLINE comprises around 19 million citations to biomedical literature, including journals and books. Medical Subject Headings (MeSH) terms are used to index MEDLINE documents and to define the class labels for the documents.

¹http://www.nlm.nih.gov/databases/databases_medline.html

As noted in Chapter 3 each associated MeSH tree code is identified within a MeSH tree structure/hierarchy and is used to identify the level in the hierarchy with which the records are associated, as well as the hierarchical relations between the classes. The MEDLINE documents were extracted using bespoke software in conjunction with the resources provided on the PubMed’s web site. PubMed² is a database that contains both MEDLINE and a number of other databases. PubMed allows users to search abstracts from MEDLINE documents by entering keywords which are then used to retrieve all the related documents (in an XML format). The extraction of secondary data from MEDLINE, to be used in relation to SAVSNET and OHSUMED, is described in the following subsections.

5.3.1.1 SAVSNET

MEDLINE was used with respect to the collection of SAVSNET documents because, although the majority of documents found in MEDLINE relate to humans, there are also a large number of documents related to animals. Apart from the related keywords, the options “English” and “animals” were selected so that the retrieved documents were in English and related to animals. A particular issue presented by the SAVSNET primary data sets was the fact that in some cases the class labels were related to very common conditions, such as “vomiting”, which allowed the retrieval of a large number of related documents. Although each data set in the SAVSNET-917-4H hierarchy was considered independently from one another in this chapter, the reason why only the first level (SAVSNET-917-1L) was considered with respect to the evaluation of the CGUSD approach was because the class labels lower down in the hierarchy are very specific, thus insufficient secondary data could be generated with respect to these classes.

5.3.1.2 OHSUMED

The OHSUMED collection is a large subset of MEDLINE, where the later comprises more recently published documents not included in the OHSUMED data set. As was described in Chapter 3, two subsets of the OHSUMED collection were considered, namely OHSUMED-CA and OHSUMED-AD, which in turn each comprised a three-level hierarchy of classes. Recall that OHSUMED-CA relates to “Cardiovascular Abnormalities” in humans, while OHSUMED-AD relates to “Animal Diseases”; thus, apart from the keywords related to the class labels and the “English” language option, the options of “humans” and “animals” were selected respectively. With regards to both OHSUMED data sets large amounts of secondary data were retrieved from MEDLINE, however there were some rare cases where fewer documents were retrieved for a certain condition, the reason was the rarity of the condition and its infrequent occurrence within the biomedical literature. Given that most of the class labels had the same number

²<http://www.ncbi.nlm.nih.gov/pubmed/>

of documents in the generated secondary data, the case of rare conditions unbalancing the data sets were not considered as having a substantial repercussion in the classification/summarisation process.

5.3.2 RCV1 (Reuters Corpus Volume 1)

The secondary data used with respect to the Reuters-21578 data collection was the RCV1 (Reuters Corpus Volume 1) data set³ which consists of 806,791 English language news stories produced by the Reuters news agency and is 37 times larger than the Reuters-21578 data set (its predecessor). The news stories contained in RCV1 were collected between 20 August 1996 and 19 August 1997 [69]. The RCV1 data set is available on request and is in XML format. Although automated coding was used to annotate each document [69], RCV1 can be considered to be a manually annotated data set because each annotation was checked by at least one human editor. Each document contained in RCV1 was annotated with three category codes as follows: (i) Topic (representing the subject area related to the document), (ii) Industry (representing the type of industry related to the document) and (iii) Region (indicating the geographical regions referred to in the document). These three main categories were subdivided into a large number of subcategories. The following subsection describes the extraction of secondary data from RCV1 to be used in relation to Reuters-21578.

5.3.2.1 Reuters-21578

For the purpose of using RCV1 as secondary data only the categories, and related subcategories, that were common with those in the Reuters-21578 data sets were used, namely: (i) Continent, (ii) Country and (iii) Type of commodity. The category “Commodity” included in the Reuters-21578-COM-2327-2L data set was not considered with respect to secondary data generation from RCV1 because it was difficult to match all the class labels found in Reuters-21578-COM-2327-2L with the topic and industry codes found in RCV1. Because the RCV1 secondary data documents were extracted according to the codes defined for the Reuters-21578 data, which were substantially fewer than the ones available in RCV1, 100,989 documents were identified. The initial number of extracted documents ($k = 100,989$) was reduced because it was so large that it could have impacted the operation of CGUSD.

5.4 Experiments and Results

This section reports on the experiments conducted to evaluate the use of the CGUSD method for summary generation. The two step classifier generation procedure that was adopted in this chapter was as follows:

³<http://about.reuters.com/researchandstandards/corpus/>

1. For each classification technique train using the secondary data and test using the primary data.
2. For each generated classifier five evaluation measures were recorded: (i) overall accuracy expressed as a percentage, (ii) Area Under the ROC Curve (AUC), (iii) precision, (iv) sensitivity/recall and (v) specificity. As in Chapter 4, of these, accuracy and AUC were considered to be the most relevant; precision, sensitivity/recall and specificity are presented so as to provide a broader insight into the effectiveness of the individual classifiers. (Each of these measures was described in Chapter 2).

The secondary data sets were generated following the criteria that was extensively described in Subsection 5.2.1. Recall that although having a large number of labelled documents to generate a classifier helps to improve its performance, it is also important to have an equal number of documents for each class. Of course an excessive amount of secondary data, balanced or not, will require a considerable computational resource.

Table 5.1 presents some statistical information concerning the primary and secondary data sets used with respect to the experiments reported in this section. From Table 5.1 it can be seen that for all the data sets considered for the evaluation of CGUSD the number of documents in the secondary data is greater than the number in the primary data (and are balanced in most cases). The reasons for not using all the data sets that were used in the evaluation presented in Chapter 4 was stated in Section 5.3. From Table 5.1 it can be noted that SAVSNET-971-3-FT and SAVSNET-917-1L have the same number of documents in the secondary data, this is because both data sets have the same class labels but related to different documents. Many interesting and different cases can be identified from Table 5.1 where the number of classes, documents in primary data and documents in secondary data vary.

As in the case of the evaluation reported in Chapter 4 the different levels in the hierarchical data sets were considered independently.

The rest of this section is divided into 12 subsections each directed at one of the data sets considered. Each subsection contains: (i) a table showing the number of documents in the primary and secondary data sets if there are no more than 20 classes (in cases where there are more than 20 classes, for space saving reasons, this information is not provided), (ii) a results table and (iii) ends with a summary of the evaluation results. For the table showing the primary and secondary data document distributions, the first column lists the class name, and the second and third columns the number of documents in the primary and secondary data per class. On the other hand, with respect to the classification results tables the first column lists the classification technique used: (i) SMO, (ii) C4.5 and (iii) RIPPER. The remaining five columns present the values obtained with respect to each of the evaluation measures used: (i) overall accuracy

Table 5.1: Statistical details for primary and secondary data sets used.

Data set	Number of classes	Number of documents	
		Primary data	Secondary data
SAVSNET-840-4-FT	4	840	1,676
SAVSNET-971-3-FT	3	971	1,287
SAVSNET-917-1L	3	917	1,287
OHSUMED-CA-3187-1L	2	3,187	4,678
OHSUMED-CA-2570-2L	16	2,570	5,534
OHSUMED-CA-834-3L	7	834	3,122
OHSUMED-AD-3393-1L	34	3,393	5,278
OHSUMED-AD-569-2L	26	569	3,926
OHSUMED-AD-292-3L	9	292	2,627
Reuters-21578-LOC-2327-1L	14	2,327	2,995
Reuters-21578-LOC-2327-2L	92	2,327	2,682
Reuters-21578-COM-2327-1L	5	2,327	3,000

expressed as a percentage (Acc), (ii) Area Under the ROC Curve (AUC), (iii) precision (Pr), (iv) sensitivity/recall (Sn/Re) and (v) specificity (Sp).

5.4.1 SAVSNET-840-4-FT

Table 5.2 presents the relation between the number of documents of the primary and secondary data sets for SAVSNET-840-4-FT with respect to each class. The primary SAVSNET-840-4-FT data comprised 840 documents unevenly distributed among four classes; the data set was very unbalanced because in the case of the “*Aggression*” class label the number of documents was much lower than the number of documents related to the other classes. Following the criteria presented in Subsection 5.2.1 the number of documents in the secondary data, extracted from MEDLINE, was 419. Thus in this case a balanced secondary data set was achieved.

Table 5.2: SAVSNET-840-4-FT primary and secondary data sets ($k = 419$).

Name of class	Number of documents	
	Primary data	Secondary data
<i>Aggression</i>	34	419
<i>Diarrhoea</i>	315	419
<i>Pruritus</i>	352	419
<i>Vomiting</i>	139	419
Total	840	1,676

From the classification results shown in Table 5.3 it can be seen that the best accuracy result was obtained using C4.5 (55.12%) and the best AUC result of 0.75 using C4.5 and RIPPER. Note that the best accuracy result is better than chance. It is conjectured that this poor performance can be attributed to the different source for the

primary data (questionnaire free text) and the secondary data (free text from medical abstracts); as well as the unbalanced nature of the primary data. The AUC results were reasonably acceptable.

Table 5.3: Classification results for the SAVSNET-840-4-FT data set.

Method	Acc (%)	AUC	Pr	Sn/Re	Sp
SMO	52.62	0.71	0.63	0.53	0.77
C4.5	55.12	0.75	0.63	0.55	0.74
RIPPER	36.19	0.75	0.84	0.36	0.95

5.4.2 SAVSNET-971-3-FT

The relationship between the number of documents in the primary and secondary data sets for SAVSNET-971-3-FT is presented in Table 5.4, from which it can be seen that the primary data comprised 971 documents unevenly divided into three classes. The primary data was very unbalanced having one class (“*Diarrhoea*”) represented by almost two thirds of the total number of documents. The secondary data, on the other hand, had a total of 1,287 documents evenly distributed among the three classes (429 documents per class).

Table 5.4: SAVSNET-971-3-FT primary and secondary data sets ($k = 429$).

Name of class	Number of documents	
	Primary data	Secondary data
<i>Diarrhoea</i>	586	429
<i>Vomiting</i>	117	429
<i>Vom & Dia</i>	268	429
Total	971	1,287

From the classification results shown in Table 5.5 it can be seen that while the best accuracy value obtained, using RIPPER, was 62.62%, the best AUC result was obtained using SMO (0.63). The results for both accuracy and AUC were better than chance but not good enough to be considered acceptable. As was stated in Chapter 3, both the “*Diarrhoea*” and the “*Vomiting*” classes are closely related to the “*Vom & Dia*” class, it can then be conjectured that the similar content of the classes might have affected the classification process despite the domain expert’s criteria and despite generating a classifier from a balanced data set. Perhaps better results would have been obtained by only considering the “*Diarrhoea*” and “*Vomiting*” classes.

Table 5.5: Classification results for the SAVSNET-971-3-FT data set.

Method	Acc (%)	AUC	Pr	Sn/Re	Sp
SMO	43.36	0.63	0.58	0.43	0.76
C4.5	30.79	0.59	0.60	0.31	0.82
RIPPER	62.62	0.60	0.61	0.63	0.55

5.4.3 SAVSNET-917

The SAVSNET-917 data set is arranged in a hierarchical manner and thus, in the context of the evaluation presented here, the levels were considered independently. Only SAVSNET-917-1L, the first level of the SAVSNET-917 hierarchy, was considered for the reasons presented in Section 5.3. The experiments carried out on SAVSNET-917-1L and the results obtained are presented in the following subsection.

5.4.3.1 SAVSNET-917-1L

Recall that the class labels of the SAVSNET-917-1L primary data set are the same ones in the SAVSNET-971-3-FT primary data set; and, although the distribution of documents per class is different, both data sets are similar in that the “*Diarrhoea*” class label has the largest number of documents (more than half of the total number of documents). As in the case of the SAVSNET-971-3-FT data set the total number of documents in the secondary data was 1,287 with 429 documents per class.

Table 5.6: SAVSNET-917-1L primary and secondary data sets ($k = 429$).

Name of class	Number of documents	
	Primary data	Secondary data
<i>Diarrhoea</i>	536	429
<i>Vomiting</i>	248	429
<i>Vom & Dia</i>	133	429
Total	917	1,287

From the classification results shown in Table 5.7 it can be seen that similar results were obtained for SAVSNET-917-1L as for SAVSNET-971-3-FT; best accuracy of 61.61% using RIPPER and best AUC of 0.62 using SMO and C4.5. The results obtained for this data set were slightly lower than the ones obtained for SAVSNET-971-3-FT in terms of both accuracy and AUC.

5.4.4 OHSUMED-CA-3187

The OHSUMED-CA-3187 data set, related to cardiovascular abnormalities, is organised in a hierarchical way where each data set in the hierarchy is considered independently

Table 5.7: Classification results for the SAVSNET-917-1L data set.

Method	Acc (%)	AUC	Pr	Sn/Re	Sp
SMO	42.31	0.62	0.54	0.42	0.74
C4.5	34.13	0.62	0.61	0.34	0.82
RIPPER	61.61	0.61	0.59	0.62	0.58

to each other. The experiments carried out with each of the data sets and the results obtained are presented in the following subsections.

5.4.4.1 OHSUMED-CA-3187-1L

Interestingly, the OHSUMED-CA-3187-1L data set comprised only two classes as shown in Table 5.8. The primary data had 3,187 documents in total and was very unbalanced, having more than two thirds of the documents allocated to the “*Congenital Heart Defects*” class. The generated secondary data set had 4,678 and was balanced, having the same number of documents related to both classes (2,339).

Table 5.8: OHSUMED-CA-3187-1L primary and secondary data sets ($k = 2,339$).

Name of class	Number of documents	
	Primary data	Secondary data
<i>Congenital Heart Defects</i>	2,339	2,339
<i>Vascular Malformations</i>	848	2,339
Total	3,187	4,678

Table 5.9 presents the obtained results: the best accuracy value was obtained with SMO (89.10%) and the best AUC value using C4.5 (0.92). Apart from having a balanced training set, another reason why these results were good is that it was a binary classification problem whereby the learning process was limited to only two classes. It is well known that SMO (an SVM style algorithm) performs well with respect to binary classification problems [112], hence the obtained results.

Table 5.9: Classification results for the OHSUMED-CA-3187-1L data set.

Method	Acc (%)	AUC	Pr	Sn/Re	Sp
SMO	89.10	0.90	0.91	0.89	0.92
C4.5	83.71	0.92	0.87	0.84	0.88
RIPPER	86.12	0.90	0.88	0.86	0.87

5.4.4.2 OHSUMED-CA-2570-2L

The OHSUMED-CA-2570-2L data set comprised 16 classes as it is shown in Table 5.10. The primary data set comprised 2,570 documents and was very unbalanced,

having some classes with hundreds of records and other classes with less than ten records. The secondary data set, on the other hand, was intended to be as balanced as possible, however there were some classes that had many fewer documents in relation to the rest of the classes: “*Crisscott Heart*” with 47 documents and “*Levocardia*” with 102 documents. It can be conjectured that the reasons for this was the rarity of these conditions, and consequently their infrequent presence in the MEDLINE database. With the exception of the aforementioned classes and the “*Scimitar Syndrome*” class, which had 367 documents, all the remaining classes had 386 related documents. The total number of documents in the secondary data set was thus 5,534, more than the number of documents in the primary data set.

Table 5.10: OHSUMED-CA-2570-2L primary and secondary data sets ($k = 386$).

Name of class	Number of documents	
	Primary data	Secondary data
<i>Cor Triatriatum</i>	21	386
<i>Coronary Vessel Anomalies</i>	218	386
<i>Crisscross Heart</i>	6	47
<i>Dextrocardia</i>	11	386
<i>Patent Ductus Arteriosus</i>	160	386
<i>Eisenmenger Complex</i>	22	386
<i>Heart Septal Defects</i>	524	386
<i>Levocardia</i>	2	102
<i>Marfan Syndrome</i>	106	386
<i>Noonan Syndrome</i>	16	386
<i>Tetralogy of Fallot</i>	153	386
<i>Aortic Coarctation</i>	237	386
<i>Transposition of Great Vessels</i>	227	386
<i>Arteriovenous Malformations</i>	525	386
<i>Scimitar Syndrome</i>	13	367
<i>Vascular Fistula</i>	329	386
Total	2,570	5,534

The obtained results, as presented in Table 5.11, were better than expected given the unbalanced nature of the primary data set and the relatively large amount of classes. The best accuracy value was obtained using RIPPER (73.91%) and the best AUC value was obtained using SMO (0.91). In this case the conjectures to explain the results were the balanced distribution of the secondary data set and the clearly different topics represented by each class; unlike in the case of the SAVSNET-971-3-FT and SAVSNET-917-1L data sets where there could have been some ambiguities with respect to the “*Diarrhoea*” and “*Vomiting*” classes and the “*Vom & Dia*” class.

Table 5.11: Classification results for the OHSUMED-CA-2570-2L data set.

Method	Acc (%)	AUC	Pr	Sn/Re	Sp
SMO	65.72	0.91	0.71	0.66	0.96
C4.5	67.03	0.85	0.70	0.67	0.96
RIPPER	73.91	0.90	0.79	0.74	0.95

5.4.4.3 OHSUMED-CA-834-3L

The OHSUMED-CA-834-3L distribution of classes for the primary and secondary data sets is shown in Table 5.12. From the table it can be seen that the primary data contained 834 documents and was very unbalanced with four classes having around 30 documents and three classes having hundreds of documents. The secondary data set, on the other hand, had a total number of 3,122 documents evenly distributed among the classes, 446 documents per class. The total number of documents contained in the secondary data set was almost four times that of the primary data set.

Table 5.12: OHSUMED-CA-834-3L primary and secondary data sets ($k = 446$).

Name of class	Number of documents	
	Primary data	Secondary data
<i>Endocardial Cushion Defects</i>	21	446
<i>Artrial Heart Septal Defects</i>	191	446
<i>Ventricular Heart Septal Defects</i>	228	446
<i>Aortopulmonary Septal Defects</i>	34	446
<i>Double Outlet Right Ventricle</i>	31	446
<i>Arterio – Arterial Fistula</i>	30	446
<i>Arteriovenous Fistula</i>	299	446
Total	834	3,122

The classification results are presented in Table 5.13. In this case, while the accuracy values were not comparable with the AUC values in terms of how good they were, they are still good considering the unbalanced nature of the primary data. Once again the AUC values gave a better insight into the performance of CGUSD with respect to the OHSUMED-CA-834-3L data set than accuracy. The number of classes did not seem to affect the classifier performance. The best accuracy value was obtained using C4.5 (66.88%) and the best AUC value was obtained with SMO (0.89).

Table 5.13: Classification results for the OHSUMED-CA-834-3L data set.

Method	Acc (%)	AUC	Pr	Sn/Re	Sp
SMO	56.46	0.89	0.76	0.57	0.96
C4.5	66.88	0.87	0.79	0.67	0.96
RIPPER	66.25	0.84	0.74	0.66	0.91

5.4.5 OHSUMED-AD-3393

OHSUMED-AD-3393 is another subset of the OHSUMED data collection, and it consists of different levels in a hierarchy in the same manner as SAVSNET-917 and OHSUMED-CA-3187. OHSUMED-AD-3393 is related to animal diseases and since there are many diseases (classes) in the first two levels of the hierarchy, only the distribution of classes of OHSUMED-AD-292-3L is presented in its respective subsection. The experiments carried out with respect to each of the data sets, and the results obtained, are presented in the following subsections.

5.4.5.1 OHSUMED-AD-3393-1L

The OHSUMED-AD-3393-1L data set comprised 3,393 documents unevenly distributed among 34 classes and the number of record per class is presented in Table A.10 of Appendix A. The primary data set was very unbalanced having many classes which were related to just a small number of documents, other classes were related to a considerably larger number of documents and one class was related to more than half of the documents in the data set. The secondary data set contains 5,278 documents which, with the exception of the “*Actinobacillosis*” class (with 64 documents), was balanced having 158 documents per class.

The classification results are presented in Table 5.14. Both the best accuracy and AUC values were obtained using SMO with 49.02% and 0.87 respectively. Overall, the accuracy values were not good, probably because of the unbalanced nature of the primary data set. The recorded AUC values, however, demonstrated that the SMO classifier performed well.

Table 5.14: Classification results for the OHSUMED-AD-3393-1L data set.

Method	Acc (%)	AUC	Pr	Sn/Re	Sp
SMO	49.02	0.87	0.77	0.49	0.97
C4.5	43.24	0.66	0.66	0.43	0.87
RIPPER	26.24	0.71	0.82	0.26	0.99

5.4.5.2 OHSUMED-AD-569-2L

OHSUMED-AD-569-2L was another data set with a considerably large amount of classes (26) and a relatively small number of documents (569). This disparity is evident in Table A.11 of Appendix A where there are classes with less than ten records and a couple of classes related to the majority of the documents. The secondary data set comprised 3,926 documents evenly distributed among all the classes (151 documents per class).

Table 5.15 presents the classification results where once again SMO had the best performance in terms of accuracy (63.33%) and AUC (0.89). RIPPER had also the best AUC value (0.89). Similarly to OHSUMED-AD-3393-1L, the AUC values were considered as good given the relatively large number of classes and the unbalanced nature of the primary data set.

Table 5.15: Classification results for the OHSUMED-AD-569-2L data set.

Method	Acc (%)	AUC	Pr	Sn/Re	Sp
SMO	63.33	0.89	0.73	0.63	0.98
C4.5	53.49	0.79	0.69	0.54	0.98
RIPPER	56.53	0.89	0.83	0.57	0.99

5.4.5.3 OHSUMED-AD-292-3L

The distribution of documents with respect to the primary and secondary data associated with the OHSUMED-AD-292-3L data set is shown in Table 5.16. The primary data comprised 292 documents and was very unbalanced having a class (“*Cryptosporidiosis*”) which related to almost half of the total number of documents. On the other hand, the secondary data contained 2,627 documents, with a balanced distribution among all the classes (292 related documents per class, with the exception of the “*Marburg Virus Disease*” class, which had 291 related documents). The total number of documents in the secondary data was almost nine times larger than that of the primary data.

Table 5.16: OHSUMED-AD-292-3L primary and secondary data sets ($k = 292$).

Name of class	Number of documents	
	Primary data	Secondary data
<i>Rift Valley Fever</i>	23	292
<i>Dirofilariasis</i>	19	292
<i>Toxocariasis</i>	23	292
<i>Babesiosis</i>	19	292
<i>Cryptosporidiosis</i>	133	292
<i>Theileriasis</i>	6	292
<i>Animal Toxoplasmosis</i>	36	292
<i>Marburg Virus Disease</i>	3	291
<i>Simian Acquired Immunodeficiency Syndrome</i>	30	292
Total	292	2,627

Table 5.17 presents the classification results. Note that good accuracy and AUC values were produced. The best accuracy value was obtained using C4.5 (85.57%) and the best AUC value using SMO (0.98). It can be conjectured that the results were good because the classes were very different from one another, aiding in the reduction of ambiguities.

Table 5.17: Classification results for the OHSUMED-AD-292-3L data set.

Method	Acc (%)	AUC	Pr	Sn/Re	Sp
SMO	82.13	0.98	0.90	0.82	0.99
C4.5	85.57	0.96	0.91	0.86	0.99
RIPPER	84.19	0.96	0.91	0.84	0.99

5.4.6 Reuters-21578-LOC-2327-2H

The Reuters-21578-LOC-2327-2H data set, as in the case of the OHSUMED and the SAVSNET-917 collections, is organised in a hierarchical way where each data set in the hierarchy is considered independently to each other. In this case all the primary data sets have the same number of documents (2,327). The experiments carried out with respect to each of these data sets, and the results obtained, are presented in the following subsections.

5.4.6.1 Reuters-21578-LOC-2327-1L

Table 5.18 presents the distribution of classes and documents with respect to the primary and secondary data sets of Reuters-21578-LOC-2327-1L. The primary data comprised 2,327 documents and was very unbalanced having the class “*America*” with nearly half of the total number of documents. The secondary data, on the other hand, consisted of 2,995 documents evenly distributed among all the classes with the exception of the “*AfricaAsia*” class, which had one document less than the rest of the classes, thus not a significant difference.

The classification results are shown in Table 5.19. Surprisingly, the accuracy results were much lower than expected. SMO produced the best performance having the highest values for both accuracy (41.30%) and AUC (0.84). It was conjectured that, despite the presence of balanced secondary data, the unbalanced nature of the primary data (as well as the number of classes) influenced the classification performance.

5.4.6.2 Reuters-21578-LOC-2327-2L

Reuters-21578-LOC-2327-2L contained 2,327 documents and was the data set used in this thesis that had the largest number of classes (92), therefore prone to be unbalanced as shown in Table A.15 of Appendix A. The classes in this data set are related to countries and were unevenly distributed. While three classes (“*Canada*”, “*UK*” and “*USA*”) were related to more than half of the total documents, there were many classes that had just one document related to them. The secondary data contained 2,682 documents evenly distributed among the majority of the classes with 33 documents per class but with 18 classes with less than 33 documents.

From Table 5.20 it can be seen that the best accuracy value was obtained using C4.5

Table 5.18: Reuters-21578-LOC-2327-1L primary and secondary data sets ($k = 214$).

Name of class	Number of documents	
	Primary data	Secondary data
<i>AfricaAmerica</i>	51	214
<i>AfricaAsia</i>	7	213
<i>Africa</i>	70	214
<i>AfricaEurope</i>	19	214
<i>AmericaAsia</i>	243	214
<i>AmericaAustralia</i>	6	214
<i>America</i>	1,021	214
<i>AmericaEurope</i>	90	214
<i>AsiaAustralia</i>	7	214
<i>Asia</i>	314	214
<i>AsiaEurope</i>	100	214
<i>Australia</i>	40	214
<i>AustraliaEurope</i>	2	214
<i>Europe</i>	357	214
Total	2,327	2,995

Table 5.19: Classification results for the Reuters-21578-LOC-2327-1L data set.

Method	Acc (%)	AUC	Pr	Sn/Re	Sp
SMO	41.30	0.84	0.66	0.41	0.95
C4.5	21.70	0.59	0.50	0.22	0.94
RIPPER	7.31	0.58	0.31	0.07	0.98

(50.62%) and the best AUC was obtained using SMO (0.88). Although the accuracy results were very low in general, the AUC results were quite satisfactory given the unbalanced nature of the primary data set, the less unbalanced nature of the secondary data set and the large number of classes.

Table 5.20: Classification results for the Reuters-21578-LOC-2327-2L data set.

Method	Acc (%)	AUC	Pr	Sn/Re	Sp
SMO	33.91	0.88	0.76	0.34	0.99
C4.5	50.62	0.73	0.62	0.51	0.95
RIPPER	28.36	0.75	0.75	0.28	1.00

5.4.7 Reuters-21578-COM-2327-2H

The Reuters-21578-COM-2327-2H data set is also organised in a hierarchical way where each data set in the hierarchy is considered independently to each other. As in the case of the Reuters-21578-LOC-2327-2H, all the primary data sets have the same number of documents (2,327). The experiments carried out with respect to this data set at the first level of the hierarchy, and the results obtained, are presented in the following subsection.

5.4.7.1 Reuters-21578-COM-2327-1L

The distribution of classes with respect to the primary and secondary data sets of Reuters-21578-COM-2327-1L is presented in Table 5.21, where it can be seen that the primary data contained 2,327 documents unevenly distributed among the classes. The secondary data, on the other hand, comprised 3,000 documents evenly distributed among all the classes. Notice that in comparison to the previous Reuters data sets used, the Reuters-21578-COM-2327-1L data set consisted of a relatively small number of classes.

Table 5.21: Reuters-21578-COM-2327-1L primary and secondary data sets ($k = 600$).

Name of class	Number of documents	
	Primary data	Secondary data
<i>Energy</i>	633	600
<i>Grains</i>	743	600
<i>Livestock</i>	105	600
<i>Metal</i>	347	600
<i>Soft</i>	499	600
Total	2,327	3,000

The classification results are presented in Table 5.22. The best accuracy value was achieved by SMO (74.95%) and the best AUC value was obtained by RIPPER (0.91).

The rest of the obtained results were also good.

Table 5.22: Classification results for the Reuters-21578-COM-2327-1L data set.

Method	Acc (%)	AUC	Pr	Sn/Re	Sp
SMO	74.95	0.90	0.78	0.75	0.94
C4.5	72.24	0.87	0.77	0.72	0.93
RIPPER	74.90	0.91	0.86	0.75	0.96

5.5 Discussion

This section presents a discussion regarding the obtained results and how they performed with respect to the objectives presented at the beginning of this chapter and with respect to the improvements made to CGUSD as first presented in [34]:

1. It was determined that text summarisation classifiers can be generated using secondary data by ensuring the compatibility between the secondary data and the primary data, and that such classifiers can be successfully applied to the primary data in order to generate summaries in the same manner as in Chapter 4. The appropriateness of the secondary data considered depends on the classes covered by the secondary data; ideally the classes featured in the secondary data should be the same as those featured in the primary data.
2. It was demonstrated that the CGUSD approach can be applied effectively with respect to free text sources different to questionnaire free text sources. For comparison purposes CGUSD was applied to text from medical abstracts (OHSUMED data sets) and to news (Reuters-21578 data sets).
3. In order to produce good quality summarisation classifiers it was necessary to attenuate the effects of having unbalanced data sets which in some cases also contained a large number of classes. The CGUSD approach allows the generation of secondary data by indicating the optimal number of documents to be retrieved in order to have a large (but manageable) and balanced secondary data set. It was demonstrated that, in most cases, it is possible to extract the required number of documents for the secondary data; however, in the uncommon case of rare classes having less documents than the other classes, it was considered not to have a substantial impact in the classification/summarisation process. Note that it is not necessarily an advantage to have balanced secondary data that contains large amounts of data, the reason being that it is expensive in terms of computational power and in the time required to process the data.

4. The operation of the three classifiers that had the best performance in Chapter 4 (SMO, C4.5 and RIPPER) with respect to CGUSD is presented in Table 5.23. Overall, the best results were obtained using SMO and RIPPER. The classification method that had the best performance for the SAVSNET data sets was RIPPER. SMO had the best performance for the OHSUMED and for the Reuters data sets. As in the case of Chapter 4, having additional evaluation measures than just only accuracy helped to have a broader insight into the results obtained, in particular by using AUC which takes into account the class priors.

Table 5.23: Best classification techniques and results.

Data set	CGUSD			Standard Classification		
	Best algorithm	Acc (%)	AUC	Best algorithm	Acc (%)	AUC
SAVSNET-840-4-FT	C4.5	55.12	0.75	SMO	89.29	0.95
SAVSNET-971-3-FT	RIPPER	62.62	0.60	SMO	74.87	0.80
SAVSNET-917-1L	RIPPER	61.61	0.61	SMO	70.34	0.75
OHSUMED-CA-3187-1L	SMO	89.10	0.90	SMO	94.84	0.93
OHSUMED-CA-2570-2L	RIPPER	73.91	0.90	SMO	80.82	0.94
OHSUMED-CA-834-3L	C4.5	66.88	0.87	SMO	78.42	0.90
OHSUMED-AD-3393-1L	SMO	49.02	0.87	SMO	82.46	0.85
OHSUMED-AD-569-2L	SMO	63.33	0.89	C4.5	76.74	0.87
OHSUMED-AD-292-3L	C4.5	85.57	0.96	C4.5	89.69	0.91
Reuters-21578-LOC-2327-1L	SMO	41.30	0.84	SMO	82.25	0.93
Reuters-21578-LOC-2327-2L	SMO	33.91	0.88	SMO	74.73	0.88
Reuters-21578-COM-2327-1L	RIPPER	74.90	0.91	SMO	95.79	0.98

Overall, using the CGUSD approach for the purpose of summary generation proved to be a viable alternative in the case where no suitable or sufficient primary data is available. The improvements made to CGUSD resulted in obtaining better results than the ones obtained in [34]. However, in general, the results were not as good as the ones obtained by applying standard classification techniques directly to the primary data as described in Chapter 4.

5.6 Summary

This chapter presented the CGUSD approach which considers the use of secondary data for the generation of classifiers where there is no suitable training data available. Secondary data related to the primary data was extracted from a related data source and then preprocessed and represented in the same way as the primary data. The three classification techniques that had the best performance in the experiments carried out in Chapter 4 were used: (i) SMO, (ii) C4.5 and (iii) RIPPER. Both the primary and secondary data had their features extracted using Chi-squared feature selection.

As in the previous chapter, the evaluation metrics used were: (i) overall accuracy expressed as a percentage, (ii) Area Under the ROC Curve (AUC), (iii) precision, (iv) sensitivity/recall and (v) specificity. The secondary data was used as the training set and the primary data as the testing set.

Overall, good classification results were obtained for most of the data sets considering that most of the primary data was unbalanced and in some cases there was a large number of classes. Hence the quality of the summaries generated was consistently good. The summaries were generated in the same manner as in the previous chapter.

Chapter 6

Using a Semi-Automated Rule Summarisation Extraction Tool (SARSET) for Text Summarisation

6.1 Introduction

The previous chapter presented a technique for generating text summarisation classifiers from the free text part of questionnaires where there was insufficient training data (or no data at all) available. This chapter presents a semi-automated classification technique called SARSET (Semi-Automated Rule Summarisation Extraction Tool). The motivation for SARSET was as follows. Previous work, presented in Chapters 4 and 5, which was directed at using standard classification techniques and secondary data, indicated that although good results could be obtained in many cases better results (in terms of classification accuracy and hence summarisation quality) might be possible. The factors that effected the operation of the previous two approaches were: (i) the different types of text data and their inherent characteristics (for example questionnaire free text that contained few words, grammatical errors, misspellings and use of specialised abbreviations and acronyms), (ii) the distribution of classes among documents because in most cases the data sets were unbalanced (although this issue was partially addressed in Chapter 5) and (iii) the operation of the techniques applied. This led the author to hypothesising that one way in which the effectiveness of the desired classification/summarisation could be improved was to involve domain experts in the classifier generation process. Hence SARSET.

The idea was to investigate a system that would allow domain experts (users) to select phrases from questionnaire returns in a training set that may be appropriate for inclusion in the antecedent of classification rules. Once these phrases had been selected it would then be possible to automatically generate variations of the suggested

phrases, using a synonym database and “wild card” characters, and produce a set of classification rules based on this collection of phrases. It would then be possible to identify and display examples from the training sets that were “covered” by these rules, and allow the user to both select appropriate rules to be included in the final classifier and to specify “exceptions” associated with particular rules. Exceptions, in this context, were specific phrases that might be covered by a rule antecedent, but which should not be used for classification purposes. This idea was realised in the form of the SARSET tool. Note that details of this tool have been previously published in [32] and [33].

The objectives of the work described in this chapter are thus as follows:

1. To determine whether the quality of the desired text summarisation classifiers can be improved by incorporating input from domain experts using a semi-automated tool (SARSET).
2. To ascertain the applicability and effectiveness of the SARSET approach when applied to free text from different sources than questionnaires.
3. To compare the operation of the SARSET approach proposed in this chapter with those of the previous chapters.

Recall that the data sets used for evaluation purposes, as noted previously in this thesis, were:

- | | |
|----------------------|-----------------------------|
| • SAVSNET-840-4-FT | • OHSUMED-CA-834-3L |
| • SAVSNET-971-3-FT | • OHSUMED-AD-3393-1L |
| • SAVSNET-917-1L | • OHSUMED-AD-569-2L |
| • SAVSNET-917-2L | • OHSUMED-AD-292-3L |
| • SAVSNET-917-3L | • Reuters-21578-LOC-2327-1L |
| • SAVSNET-917-4L | • Reuters-21578-LOC-2327-2L |
| • OHSUMED-CA-3187-1L | • Reuters-21578-COM-2327-1L |
| • OHSUMED-CA-2570-2L | • Reuters-21578-COM-2327-2L |

With respect to the evaluation presented later in this chapter the SAVSNET-840-4-TD+FT and SAVSNET-971-3-TD+FT data sets were not used because these data sets included tabular data (the proposed SARSET approach does not lend itself to application to tabular data). Also it had been previously established (as reported in

Chapter 4) that usage of tabular data did not benefit the classification summarisation process. Thus 16 data sets were used to evaluate SARSET.

The remainder of this chapter is arranged as follows. Section 6.2 describes the SARSET methodology in detail as well as its operation. A comprehensive description of the experiments carried out on the data sets, as well as an interpretation of the obtained results is then presented in Section 6.3, and a discussion of how the proposed technique performed is presented in Section 6.4. Finally, a summary of the chapter is presented in Section 6.5.

6.2 The SARSET Methodology

The proposed SARSET methodology is presented in this section, which is divided into two subsections: Subsection 6.2.1 presents the formal definition of the problem that this technique addresses, and Subsection 6.2.2 describes the implementation of the proposed technique. Note that the text summarisation element of the process is carried out in the same way as described in Chapter 4, and is thus not commented on further in this chapter.

6.2.1 Problem definition

SARSET is specifically directed at the summarisation of questionnaire returns where the quantity and the quality of the text is limited; although, as will be demonstrated later in this chapter, it can equally well be used for other forms of text. The expected input is thus a collection of n questionnaires, $Q = \{q_1, q_2, \dots, q_n\}$, where each questionnaire comprises a tabular component and a free text component, $q_i = \{T_i, D_i\}$ (where i is a numeric questionnaire identifier). The text element contains sequences of words, numbers, punctuation and other printable characters. We indicate the set of free text components as $D = \{D_1, D_2, \dots, D_m\}$. The objective is then to summarise the free text element of the questionnaires by searching for patterns in the document set that lead to particular classifications according to a given class set $C = \{C_1, C_2, \dots, C_n\}$. Note that we indicate the complete set of class labels using the identifier C . Each class in C has a set of class values associated with it, $C_i = \{c_{i_1}, c_{i_2}, \dots, c_{i_k}\}$ (where k is the number of values). Thus we have a multi-class problem. Given a pattern (phrase) s , that might indicate a class value c_{i_j} , this can be expressed in the form of a classification rule $s \Rightarrow c_{i_j}$. The idea is that we create a collection R of rule based classifiers $R = \{R_1, R_2, \dots, R_n\}$, one classifier per class, and that this collection of classifiers can then be applied to classify (and hence summarise) a questionnaire collection. Note that the two previously described approaches could also be used to address the multi-class problem by generating a number of individual classifiers, one per class. The overall objective is thus to relate the input $Q = \{q_1, q_2, \dots, q_n\}$ to a set of class labels

$\{\{c_{1_1}, c_{1_2}, \dots, c_{1_n}\}, \{c_{2_1}, c_{2_2}, \dots, c_{2_n}\}, \dots, \{c_{n_1}, c_{n_2}, \dots, c_{n_n}\}\}$ such that a set of labels $\{c_{i_1}, c_{i_2}, \dots, c_{i_n}\}$ can be associated with each questionnaire q_i which can then be used to generate the desired summary for that questionnaire free text.

6.2.2 Classifier Generation Using SARSET (Semi-Automated Rule Summarisation Extraction Tool)

In this section the SARSET methodology is described in more detail. SARSET comprises 5 steps as shown in Figure 6.1 (each is considered in the following subsections). Broadly the SARSET process can be described as follows:

1. The user identifies a relevant phrase in the free text questionnaire data and the system then automatically identifies variations of this phrase to give a set of phrases P .
2. The system extracts the subset of questionnaires in D that feature (are “covered” by) the phrases in P .
3. If a suitable phrase p_i can be identified in P (one that serves to identify a class value c_i): (i) generate a classification rule with p_i as the antecedent and c_i as the consequent, and add to R , (ii) if necessary add exceptions to the *exceptions base*, (iii) remove p_i from P . Otherwise go to 5.
4. Repeat 3.
5. Exit if a suitably effective classifier has been generated. Otherwise go to 1.

Note that prior to commencement of the process the “documents” in D are preprocessed so that numbers and symbols are removed, but keeping phrase delimiters such as commas, semicolons and full stops in order to have a clean but coherent free text from which the domain experts can identify relevant phrases.

6.2.3 Phrase identification and generation of phrase variations (Step 1)

The first step in the SARSET process, as indicated above, involves the participation of a domain expert (or user). In the Graphical User Interface (GUI) provided by the SARSET tool (Figure 6.2), the first document $d_i \in D$ is presented to the user, who then identifies a phrase relevant to the application domain which describes the document in terms of some summary class type. The user identified phrase is conceptualised as an ordered sequence of k ($1 \leq k \leq 5$) words that includes at least one keyword as determined by the user, and one or more non-keywords (punctuation is ignored). Phrase variations are then generated for the identified phrase, whereby non-keywords

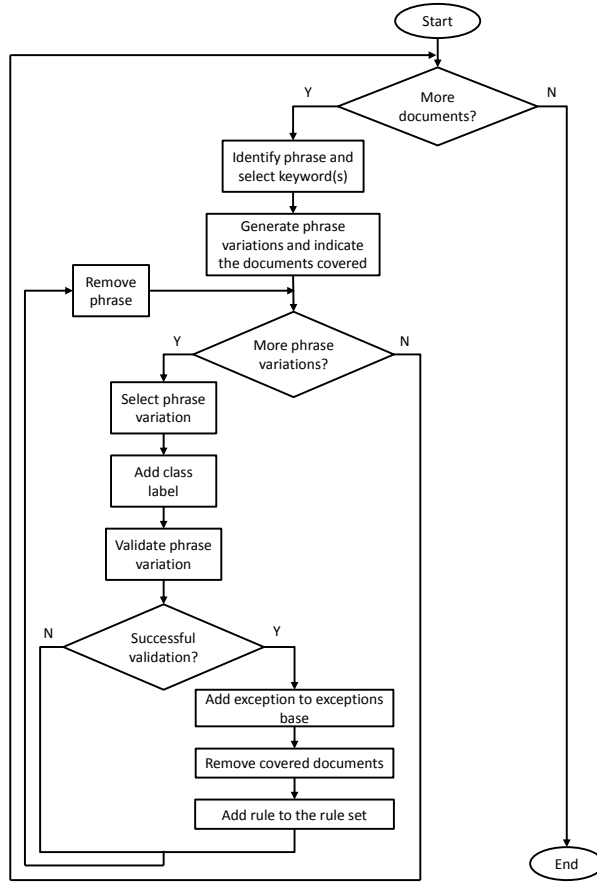


Figure 6.1: The SARSET methodology.

are replaced with “wild card markers” which can be matched to any word in the document set using a one-to-one matching. The idea here is that given a likely phrase that may become part of a classification rule, similar phrases sharing the same pattern may also be useful. For the phrase variation construction synonyms for the keyword(s) selected are also considered in order to broaden the coverage of the phrase pattern and its related phrases. The synonyms used are identified automatically using a Lucene¹ index that contains the synonyms defined in the WordNet² database.

For example, suppose the phrase “continue with bland diet” has been identified and suppose the set of keywords is $K = \{diet\}$. Consequently the set of non-keywords is $W = \{continue, with, bland\}$. SARSET automatically builds all the variations of this phrase. In this case, including synonyms, we get:

$$\begin{aligned}
 P = \{ & \{continue, with, bland, diet\}, \{?, with, bland, diet\}, \\
 & \{continue, ?, bland, diet\}, \{continue, with, ?, diet\}, \{?, ?, bland, diet\}, \\
 & \{?, with, ?, diet\}, \{continue, ?, ?, diet\}, \{?, ?, ?, diet\}, \{?, bland, diet\}, \\
 & \{continue, ?, diet\}, \{?, diet\}, \{continue, with, bland, dieting\},
 \end{aligned}$$

¹<http://lucene.apache.org/core/>

²<http://wordnet.princeton.edu/>

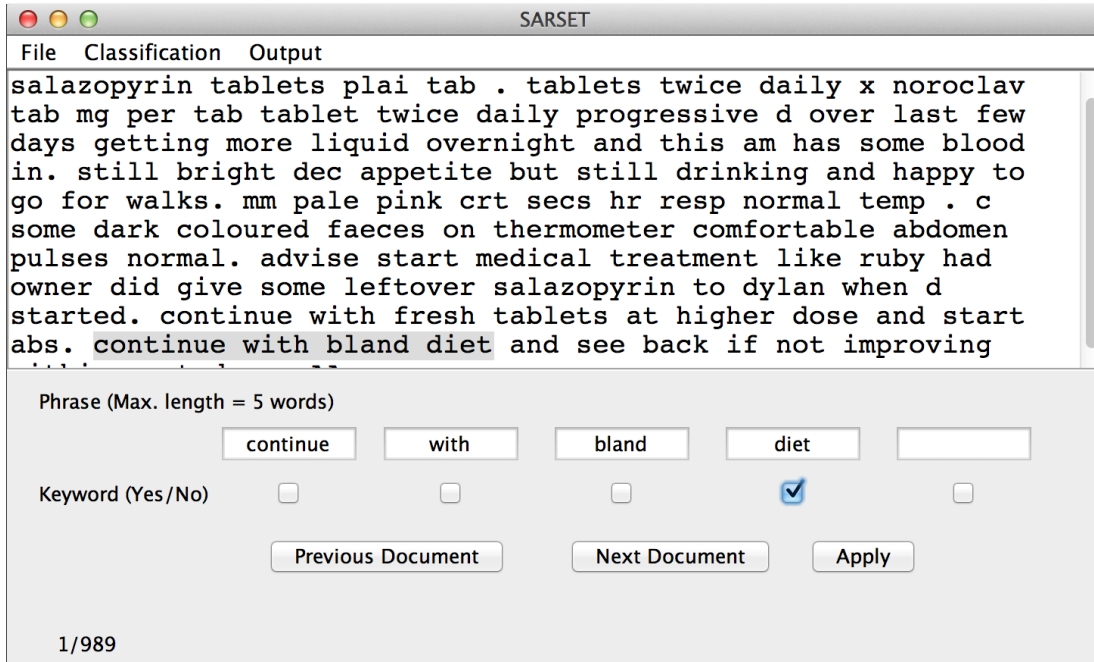


Figure 6.2: Main window of SARSET.

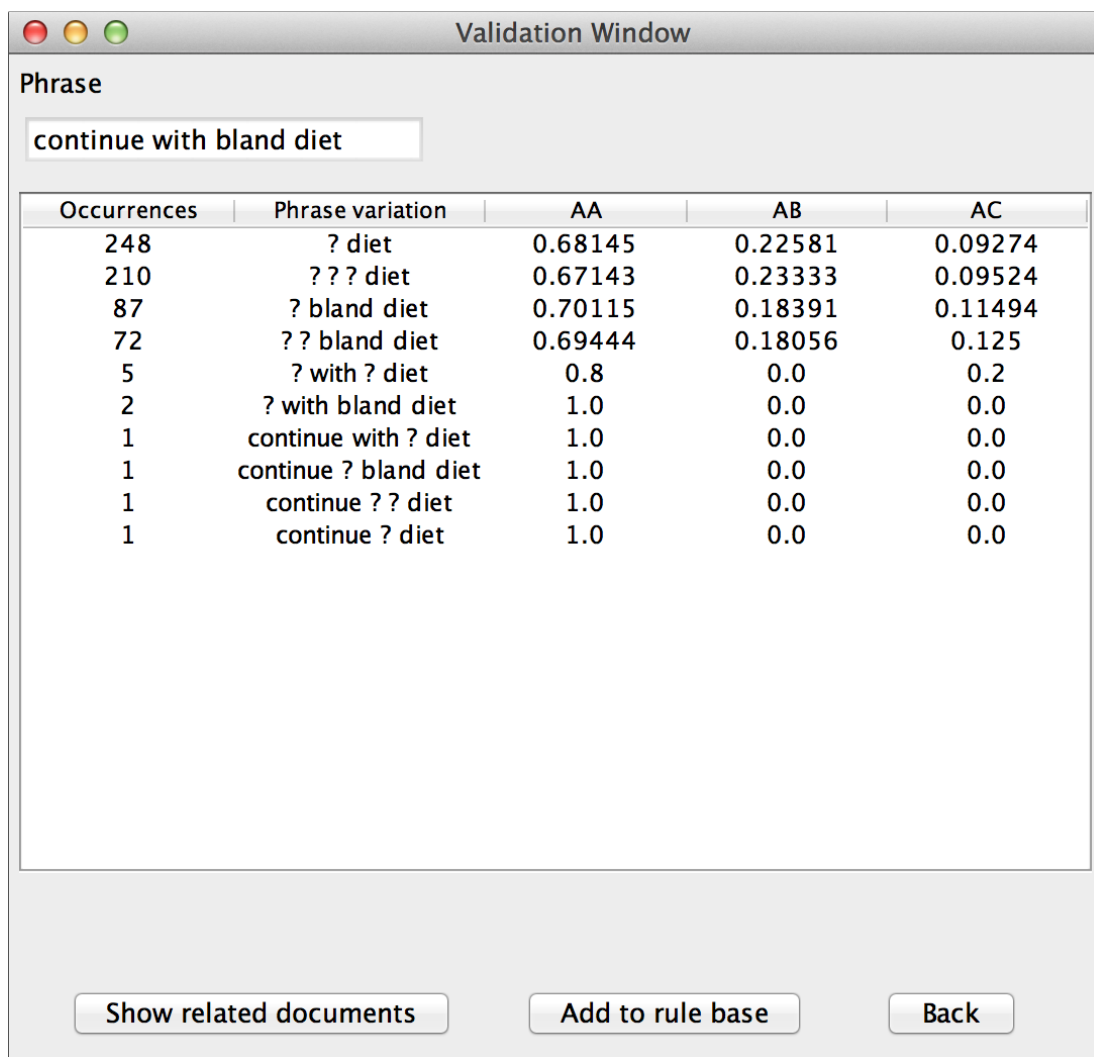
$$\begin{aligned} &\{?, with, bland, dieting\}, \{continue, ?, bland, dieting\}, \\ &\{continue, with, ?, dieting\}, \{?, ?, bland, dieting\}, \{?, with, ?, dieting\}, \\ &\{continue, ?, ?, dieting\}, \{?, ?, ?, dieting\}, \{?, bland, dieting\}, \\ &\{continue, ?, dieting\}, \{?, dieting\} \end{aligned}$$

($|P| = 22$). In the above the first phrase is the phrase identified by the user, the following 10 are variations identified by the system, and the following 11 are the original phrase and its variations using “dieting” as a synonym for “diet”.

6.2.4 Identification of questionnaires covered by identified phrases (Step 2)

The second step is the automatic retrieval of documents in D which are covered by the identified phrase in the set P . SARSET presents a list of the phrases ordered according to the frequency with which they appear in D , and gives the probabilities with which each phrase is associated with each class (see Figure 6.3). Note that in this example the phrase used is “continue with bland diet” to which SARSET generated variations including “dieting” as a synonym for “diet”, however the phrase variations which included “dieting” did not match any phrase in the documents in D . Recall that in some cases the names of the classes are too long or can comprise more than two words; taking this into account, and in order to allow for the visualisation and handling of data sets with large number of classes within SARSET, instead of presenting the names of the classes two-letter codes related to each of the classes are presented instead. When

the actual summaries are generated, the codes are replaced by their corresponding class names.



The Validation Window of SARSET displays a table of phrase variations and their associated probabilities for three classes: AA, AB, and AC. The table is titled 'Validation Window' and has a 'Phrase' input field at the top. The input field contains the text 'continue with bland diet'. Below the input field is a table with five columns: Occurrences, Phrase variation, AA, AB, and AC. The table lists 11 phrase variations and their corresponding probabilities. At the bottom of the window are three buttons: 'Show related documents', 'Add to rule base', and 'Back'.

Occurrences	Phrase variation	AA	AB	AC
248	? diet	0.68145	0.22581	0.09274
210	??? diet	0.67143	0.23333	0.09524
87	? bland diet	0.70115	0.18391	0.11494
72	?? bland diet	0.69444	0.18056	0.125
5	? with ? diet	0.8	0.0	0.2
2	? with bland diet	1.0	0.0	0.0
1	continue with ? diet	1.0	0.0	0.0
1	continue ? bland diet	1.0	0.0	0.0
1	continue ?? diet	1.0	0.0	0.0
1	continue ? diet	1.0	0.0	0.0

Figure 6.3: Validation window of SARSET.

The frequency and the probabilities associated with each phrase allows the user to determine their relevance with respect to the classification task. If desired, the user can inspect the documents covered by each phrase to see the context in which the phrase appears (for example so as to identify potential exceptions) (see Figure 6.4).

6.2.5 Rule generation (Steps 3 and 4)

In steps 3 and 4 the user selects suitable phrases to be included in classification rules. This process continues until no more phrases can be identified. For each identified phrase a rule is constructed and added to the rule set R so far. An example of a generated rule is: $? \text{ bland diet} \Rightarrow AA$, in which the antecedent of the rule consists of a phrase variation and the consequent the name of a class (in this case a code which

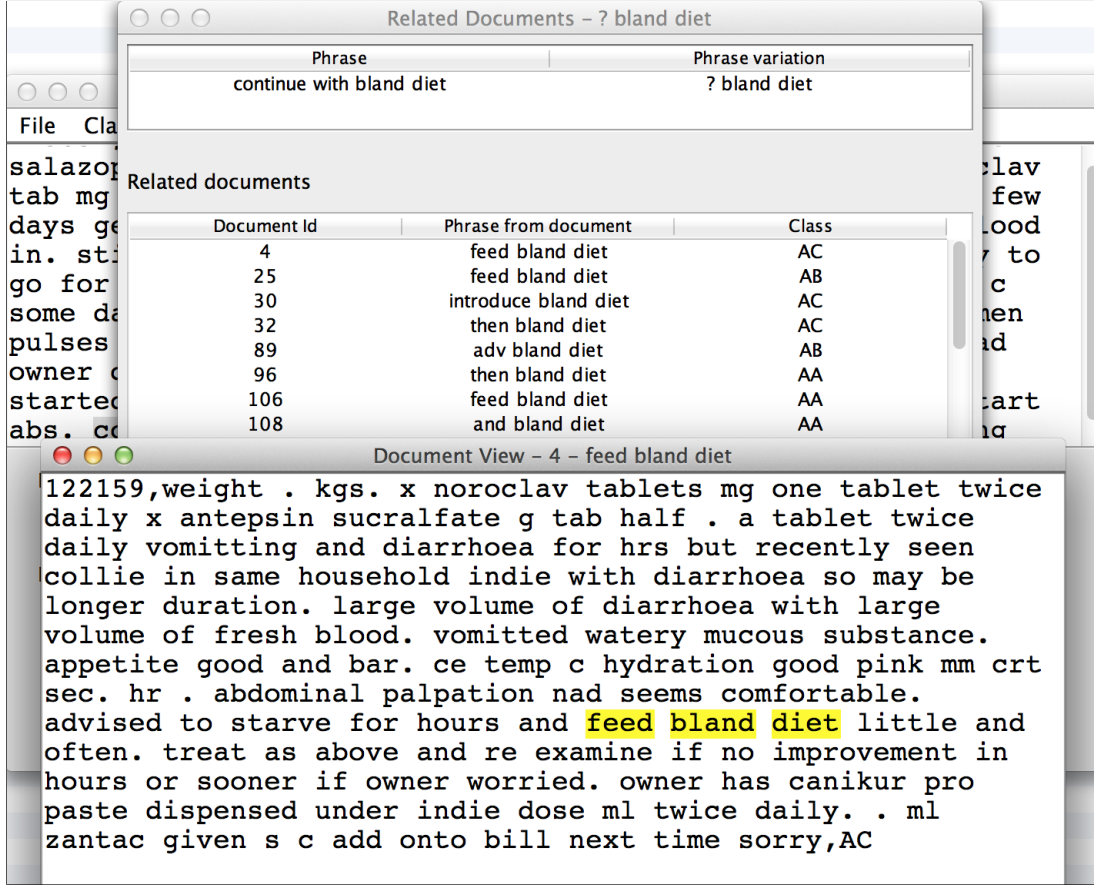


Figure 6.4: Inspection of documents.

corresponds to the class *Diarrhoea*). The appropriateness of phrases is judged by their associated frequency count and probability of being associated with a particular class. If, once a rule is generated, the user can identify exceptions these are included in an exceptions base. As noted earlier in this chapter exceptions are phrase patterns that should not be “covered” by particular identified rules. For example given the rule $? \text{ bland diet} \Rightarrow AA$ we might not want the antecedent to cover phrases such as *aspirin bland diet* and *without bland diet*, in which case *aspirin bland diet* and *without bland diet* would be added to the exceptions base. Documents that are covered by generated classification rules are not removed from the document set. The argument for not removing documents in the SARSET approach is that the free text element of questionnaires may contain more than one phrase that can be considered relevant.

The generated classification rules are ranked according to the coverage given by the probability of a phrase variation being associated with a certain class, in other words by their support, which is calculated as follows:

$$supp(A \Rightarrow B) = \frac{||A \wedge B||}{||D||} \quad (6.1)$$

Where A is the antecedent of the rule (phrase variation), B is the consequent (name of the associated class) and D is the document set. Support is thus the proportion of the number of transactions covering a rule $A \Rightarrow B$ with respect to the total number of transactions in the given data set. Recall that the domain expert considers the frequency of the phrase variation within the document set D as well as its relevance with respect to the application domain, when selecting phrase variations that will become part of the antecedent of a rule. Note also that high ranked rules are “fired” before other rules. Rules were also arranged according to their antecedent size so as to facilitate effective “look-up” (see Subsection 6.2.7 below for more detail on the rational for this).

6.2.6 Continuation of the process or exit (Step 5)

The overall process continues until a suitably effective classifier is arrived at. This will be a decision for the domain expert (user). However, it is suggested that this should be when all documents in the training set are covered by at least one rule in R or no more rules can be generated.

6.2.7 Applying classification rules to unseen documents

Once a classifier has been generated it may be applied to summarise unseen questionnaire data (or other types of free text). In practice several classifiers will be produced to cover each of the classes included in the questionnaire set (the set C identified in the problem definition). To apply the classifiers generated using the SARSET methodology, the document collection to which the classifiers are to be applied must first be preprocessed in the same manner as the document collection that was used to generate the rules, that is, by removing numbers and symbols and by keeping phrase delimiters (commas, semicolons and full stops). A collection of feature vectors, as used with respect to some text classification systems, was not produced because it was not necessary and because it would have been computationally expensive to generate. Phrases of k ($1 \leq k \leq 5$) words size are then identified in the documents and the classification rules applied according to their ranking and antecedent size (thus rules whose antecedent comprises 2-words are applied to 2-words phrases). Phrases from an unclassified document that match the antecedent (a phrase pattern) of a classification rule are classified according to the rule that is “fired” first. In the case where a document cannot be classified because there is no phrase pattern that matches any of the phrases in the document a default class is selected (the class that appeared most frequently in the training set).

6.3 Experiments and Results

This section reports on the experiments conducted to evaluate the operation of SARSET using data sets containing different types of text. A 50:50 training-test set split was adopted. Evaluation was conducted by applying the classifiers to the test set, the original training sets and the entire data set (training and test set). The reason was that since the classifiers were generated based on the relevant phrases that were identified in the text and their variations, in most cases not all the documents in the document set were covered, as opposed to the typical case when generating classifiers using standard techniques such as decision trees. Evaluating the generated classifiers on the entire document set and on the first and second halves of it ensured that all the documents were used for the evaluation. TCV was not used because of the resource intensive nature of the SARSET approach. As in previous chapters, five evaluation measures were used: (i) overall accuracy expressed as a percentage, (ii) Area Under the ROC Curve (AUC), (iii) precision, (iv) sensitivity/recall and (v) specificity. Again, accuracy and AUC were considered to be the most relevant evaluation measures; as in the case of the evaluation results presented in previous chapters, precision, sensitivity/recall and specificity were recorded so as to provide a broader insight into the effectiveness of the individual classifiers. (Each of these measures was described in Chapter 2.)

The SARSET approach requires intervention from a domain expert (user) who has to use his/her expertise with respect to the domain of the document set in order to identify and select the most relevant phrases with respect to a certain class. For experimental purposes it was not possible to enlist the full time services of a domain expert, instead the author received instruction from two domain experts so that he could act as a “domain expert” himself. The author was also able to utilise his familiarity with the SAVSNET questionnaire collection obtained during the time working on the PhD programme of research.

With respect to the OHSUMED collection recall that the OHSUMED-CA data set was related to “Cardiovascular Abnormalities” in humans and the OHSUMED-AD data set was related to “Animal Diseases”. Although the latter data set was related to animal diseases more class labels had to be taken into account than in the case of the SAVSNET collection. To aid the author in selecting relevant phrases from the medical and the veterinary science areas three strategies were followed:

1. *Identify relevant phrases in the text of the documents.* This first strategy was part of the SARSET operation and in this case it was taken into account that the author was neither a medical doctor nor a veterinary. However, considering previous advice from domain experts and using common sense, relevant phrases were identified based on the names of diseases and drugs.
2. *Use the names of the classes as part of relevant phrases.* Unlike the classes in

the SAVSNET data sets, in most cases the classes in the OHSUMED data sets were related to more specific topics, allowing for the identification of phrases that clearly made a distinction between the related classes. In other words, less ambiguous rules could be generated.

3. *Use lists of diseases and drugs from medical and veterinary web pages that were related to the class topics.* The lists used provided more information (names of diseases and drugs) to aid the identification of relevant phrases in the text documents. This also allowed the author to better understand the relationship between words in the documents and their respective classes.

Regarding the Reuters-21578 collection, the author was able to become familiar with the topics addressed in the text documents (news stories) while generating the data sets that were used for evaluation purposes in this research based on the original Reuters-21578 collection. Some additional research was conducted by the author regarding “commodities”, which are what the classes in Reuters-21578-COM-2327-2L relate to, having gained this understanding the identification of relevant phrases and keywords in the Reuters-21578 data sets was a straightforward process. The names of the classes were also used to identify relevant phrases in the free text.

Thus the classifiers were generated according to the best knowledge possessed by the author regarding the data sets used. In a similar manner to the arrangement in Chapters 4 and 5 the rest of this section is divided into 16 subsections each directed at one of the data sets considered. Each subsection contains a results table where the first column lists which part of the data set was used for testing purposes: (i) entire data set, (ii) first half and (iii) second half. The remaining five columns present the values obtained with respect to each of the evaluation measures used: (i) overall accuracy expressed as a percentage (Acc), (ii) Area Under the ROC Curve (AUC), (iii) precision (Pr), (iv) sensitivity/recall (Sn/Re) and (v) specificity (Sp).

6.3.1 SAVSNET-840-4-FT

Table 6.1 presents the results obtained using SARSET on the SAVSNET-840-4-FT data set, which comprised 840 documents unevenly distributed among 4 classes. The best accuracy result (88.86%) was obtained when SARSET was applied to the second half of the data set. The best AUC result (0.82) was obtained when SARSET was applied to the entire data set, and to the first and second halves. As was to be expected, the results from applying SARSET to the entire data set and the first half were very similar to those obtained when SARSET was applied to the second half. The results were good considering the unbalanced nature of the SAVSNET-840-4-FT data set.

Table 6.1: Classification results for the SAVSNET-840-4-FT data set.

	Acc (%)	AUC	Pr	Sn/Re	Sp
Entire data set	88.02	0.82	0.82	0.93	0.84
First half	87.65	0.82	0.81	0.93	0.84
Second half	88.86	0.82	0.83	0.94	0.85

6.3.2 SAVSNET-971-3-FT

The classification results for the SAVSNET-971-3-FT data set are shown in Table 6.2. SAVSNET-971-3-FT consisted of 971 documents unevenly distributed among 3 classes. The best accuracy and AUC results (76.13% and 0.82 respectively) were obtained when SARSET was applied to the first half of the data set. As in the experiments presented in previous chapters with the SAVSNET-971-3-FT data set it can be conjectured that having a class (*“Vom & Dia”*) closely related to other classes (*“Diarrhoea”* and *“Vomiting”*) may have caused ambiguities during the classification process.

Table 6.2: Classification results for the SAVSNET-971-3-FT data set.

	Acc (%)	AUC	Pr	Sn/Re	Sp
Entire data set	75.23	0.79	0.67	0.80	0.72
First half	76.13	0.82	0.68	0.81	0.73
Second half	75.10	0.78	0.67	0.80	0.72

6.3.3 SAVSNET-917

As mentioned in previous chapters, the SAVSNET-917 data set is arranged in a hierarchical manner having four levels. SARSET was applied to each level in the SAVSNET-917 hierarchy considering each level independently with respect to each other. The results for each level are presented in the following subsections.

6.3.3.1 SAVSNET-917-1L

The classification results for the SAVSNET-917-1L data set are shown in Table 6.3. SAVSNET-917-1L comprised 917 documents unevenly distributed among 3 classes. The best accuracy and AUC results (74.36% and 0.78 respectively) were obtained when SARSET was applied to the second half of the data set. Surprisingly, the results obtained were lower than the ones obtained for SAVSNET-971-3-FT, which has the same classes and is just as unbalanced. Both SAVSNET-917-1L and SAVSNET-971-3-FT vary in their number of documents and in their distribution with respect to the classes. As in the case of SAVSNET-971-3-FT it can also be conjectured that having a class (*“Vom & Dia”*) closely related to other classes (*“Diarrhoea”* and *“Vomiting”*) may have caused ambiguities during the classifier generation process.

Table 6.3: Classification results for the SAVSNET-917-1L data set.

	Acc (%)	AUC	Pr	Sn/Re	Sp
Entire data set	72.78	0.77	0.64	0.78	0.69
First half	72.41	0.77	0.63	0.77	0.69
Second half	74.36	0.78	0.66	0.79	0.71

6.3.3.2 SAVSNET-917-2L

Table 6.4 presents the results obtained using SARSET on the SAVSNET-917-2L data set, which contained 917 documents unevenly distributed among 3 classes. The best accuracy and AUC results (75.42% and 0.59 respectively) were obtained when SARSET was applied to the first half of the data set. In terms of AUC the results were significantly lower than those obtained with the SAVSNET-840-4-FT, SAVSNET-971-3-FT and SAVSNET-917-1L data sets. The reasons could have been: (i) having a very unbalanced distribution of the documents among the classes with one class having 65.87% of the documents, and (ii) the possible ambiguity of the classes with respect to the free text in the documents.

Table 6.4: Classification results for the SAVSNET-917-2L data set.

	Acc (%)	AUC	Pr	Sn/Re	Sp
Entire data set	73.38	0.56	0.65	0.78	0.70
First half	75.42	0.59	0.67	0.80	0.72
Second half	72.29	0.54	0.63	0.77	0.69

6.3.3.3 SAVSNET-917-3L

The classification results for the SAVSNET-917-3L data set are shown in Table 6.5. SAVSNET-917-3L comprised 917 documents unevenly distributed among 3 classes. The best accuracy and AUC results (71.66% and 0.74 respectively) were obtained when SARSET was applied to the second half of the data set. The results were considered to be good, despite a very unbalanced data set and difficulties encountered in relating documents to classes.

Table 6.5: Classification results for the SAVSNET-917-3L data set.

	Acc (%)	AUC	Pr	Sn/Re	Sp
Entire data set	70.73	0.73	0.61	0.76	0.67
First half	70.98	0.73	0.61	0.76	0.68
Second half	71.66	0.74	0.62	0.77	0.68

6.3.3.4 SAVSNET-917-4L

The classification results for the SAVSNET-917-4L data set are shown in Table 6.6. SAVSNET-917-4L consisted of 917 documents unevenly distributed among 5 classes. The best accuracy and AUC results (52.97% and 0.07 respectively) were obtained when SARSET was applied to the entire data set. The results obtained were very disappointing because, while in terms of accuracy they were better than chance, in terms of AUC they were extremely low. It was conjectured that the reasons for these poor results were: (i) the unbalanced nature of the data set and (ii) the lack of information related to the classes in the document set.

Table 6.6: Classification results for the SAVSNET-917-4L data set.

	Acc (%)	AUC	Pr	Sn/Re	Sp
Entire data set	52.97	0.07	0.08	0.79	0.52
First half	52.09	0.05	0.06	0.79	0.51
Second half	52.28	0.06	0.06	0.78	0.51

6.3.4 OHSUMED-CA-3187

The OHSUMED-CA-3187 data set, related to cardiovascular abnormalities, is organised in a hierarchical manner and thus, as in the case of previously considered hierarchical data sets, each data set in the hierarchy was considered independently. The experiments carried out with respect to each of the data sets, and the results obtained, are presented in the following subsections.

6.3.4.1 OHSUMED-CA-3187-1L

Table 6.7 presents the results obtained when applying SARSET to the OHSUMED-CA-3187-1L data set, which comprised 3,187 documents unevenly distributed among 2 classes. The best accuracy and AUC results (77.21% and 0.87 respectively) were obtained when SARSET was applied to the second half of the data set. The obtained results were good despite having a very unbalanced data set where one class was associated with more than 70% of the documents.

Table 6.7: Classification results for the OHSUMED-CA-3187-1L data set.

	Acc (%)	AUC	Pr	Sn/Re	Sp
Entire data set	76.37	0.86	0.76	0.76	0.76
First half	75.53	0.86	0.76	0.76	0.76
Second half	77.21	0.87	0.77	0.77	0.77

6.3.4.2 OHSUMED-CA-2570-2L

The classification results for the OHSUMED-CA-2570-2L data set are shown in Table 6.8. OHSUMED-CA-2570-2L contained 2,570 documents unevenly distributed among 16 classes. The best accuracy and AUC results (67.78% and 0.35 respectively) were obtained when SARSET was applied to the first half of the data set. From the results it can be seen that, although the classes of the OHSUMED-CA-2570-2L data set were very different from one another thus reducing the possible ambiguities, the results in terms of AUC were very low. It can be conjectured that the unbalanced distribution of documents among the classes may have been the reason for having such low AUC results. The AUC results contrasted with the accuracy results, which appeared to be affected by the distribution of the documents with respect to the classes.

Table 6.8: Classification results for the OHSUMED-CA-2570-2L data set.

	Acc (%)	AUC	Pr	Sn/Re	Sp
Entire data set	67.02	0.33	0.38	0.90	0.61
First half	67.78	0.35	0.40	0.91	0.61
Second half	66.23	0.32	0.37	0.90	0.60

6.3.4.3 OHSUMED-CA-834-3L

Table 6.9 presents the results obtained using SARSET on the OHSUMED-CA-834-3L data set, which contained 834 documents unevenly distributed among 7 classes. The best accuracy and AUC results (82.19% and 0.61 respectively) were obtained when SARSET was applied to the second half of the data set. Similarly to the OHSUMED-CA-2570-2L data set, the AUC results helped to understand how well SARSET performed on the OHSUMED-CA-834-3L data set despite obtaining accuracy values above 80%.

Table 6.9: Classification results for the OHSUMED-CA-834-3L data set.

	Acc (%)	AUC	Pr	Sn/Re	Sp
Entire data set	81.67	0.59	0.69	0.93	0.75
First half	81.16	0.58	0.68	0.93	0.75
Second half	82.19	0.61	0.69	0.93	0.76

6.3.5 OHSUMED-AD-3393

The OHSUMED-AD-3393 data set is related to animal diseases and is organised in a hierarchical way in the same manner as SAVSNET-917 and OHSUMED-CA-3187. The experiments carried out with respect to each of the data sets, and the results obtained, are presented in the following subsections.

6.3.5.1 OHSUMED-AD-3393-1L

Table 6.10 presents the results obtained when SARSET was applied to the OHSUMED-AD-3393-1L data set, which consisted of 3,393 documents unevenly distributed among 34 classes. The best accuracy and AUC results (84.55% and 0.65 respectively) were obtained when SARSET was applied to the second half of the data set. Despite obtaining good accuracy results, as in previous cases, the AUC results indicated that the outcome was not as good as first indicated by the accuracy results. It was conjectured that the reasons for this are: (i) the unbalanced nature of the OHSUMED-AD-3393-1L data set, and (ii) the lack of examples (phrases) related to certain classes within the document set.

Table 6.10: Classification results for the OHSUMED-AD-3393-1L data set.

	Acc (%)	AUC	Pr	Sn/Re	Sp
Entire data set	84.27	0.64	0.70	0.99	0.76
First half	83.99	0.63	0.69	0.99	0.76
Second half	84.55	0.65	0.70	0.99	0.77

6.3.5.2 OHSUMED-AD-569-2L

Table 6.11 presents the results obtained using SARSET on the OHSUMED-AD-569-2L data set, which contained 569 documents unevenly distributed among 26 classes. The best accuracy and AUC results (78.84% and 0.57 respectively) were obtained when SARSET was applied to the first half of the data set. Both the accuracy and AUC results were relatively poor. As in the case of OHSUMED-AD-3393-1L data set, it can be conjectured that the reasons for these results were: (i) the unbalanced nature of the OHSUMED-AD-569-2L data set, and (ii) the lack of examples (phrases) related to certain classes within the document set.

Table 6.11: Classification results for the OHSUMED-AD-569-2L data set.

	Acc (%)	AUC	Pr	Sn/Re	Sp
Entire data set	78.71	0.55	0.59	0.97	0.71
First half	78.84	0.57	0.59	0.97	0.71
Second half	78.58	0.52	0.59	0.97	0.70

6.3.5.3 OHSUMED-AD-292-3L

The classification results for the OHSUMED-AD-292-3L data set are shown in Table 6.12. OHSUMED-AD-292-3L comprised 292 documents unevenly distributed among 9 classes. The best accuracy result (83.32%) was obtained when SARSET was applied

to the second half of the data set, while the best AUC result (0.75) was obtained when SARSET was applied to the first and to the second halves of the data set. Both the accuracy and AUC results were considered to be acceptable given the unbalanced distribution of documents among classes. Note that OHSUMED-AD-292-3L contains many fewer documents than the other OHSUMED data sets.

Table 6.12: Classification results for the OHSUMED-AD-292-3L data set.

	Acc (%)	AUC	Pr	Sn/Re	Sp
Entire data set	82.08	0.75	0.68	0.94	0.75
First half	80.87	0.74	0.66	0.94	0.74
Second half	83.32	0.75	0.70	0.95	0.76

6.3.6 Reuters-21578-LOC-2327-2H

The Reuters-21578-LOC-2327-2H data set, as in the case of the OHSUMED data sets and the SAVSNET-917 data set, is organised in a hierarchical way such that each data set in the hierarchy was considered independently. The experiments carried out with respect to each of the data sets and the results obtained are presented in the following subsections.

6.3.6.1 Reuters-21578-LOC-2327-1L

Table 6.13 presents the results when applying SARSET to the Reuters-21578-LOC-2327-1L data set, which contained 2,327 documents unevenly distributed among 14 classes. The best accuracy result (79.88%) was obtained when SARSET was applied to the first half of the data set, and the best AUC result (0.60) was obtained when SARSET was applied to the entire data set and to the second half of the data set.

Table 6.13: Classification results for the Reuters-21578-LOC-2327-1L data set.

	Acc (%)	AUC	Pr	Sn/Re	Sp
Entire data set	79.75	0.60	0.62	0.96	0.72
First half	79.88	0.59	0.63	0.96	0.72
Second half	79.58	0.60	0.62	0.96	0.72

6.3.6.2 Reuters-21578-LOC-2327-2L

Table 6.14 presents the results obtained using SARSET with respect to the Reuters-21578-LOC-2327-2L data set, which contained 2,327 documents unevenly distributed over 92 classes. The best accuracy result (77.24%) was obtained when SARSET was applied to the first half of the data set. The best AUC result (0.52) was obtained when SARSET was applied to the entire data set and to the first half of the data set. As

was expected, because of the large number of classes, the results in terms of AUC were just slightly better than chance. In terms of accuracy the results were better, but this could have been as a result of having an unbalanced distribution of documents among a large number of classes.

Table 6.14: Classification results for the Reuters-21578-LOC-2327-2L data set.

	Acc (%)	AUC	Pr	Sn/Re	Sp
Entire data set	76.65	0.52	0.53	0.99	0.67
First half	77.24	0.52	0.55	0.99	0.69
Second half	75.14	0.51	0.51	0.99	0.67

6.3.7 Reuters-21578-COM-2327-2H

The Reuters-21578-COM-2327-2H data set is also organised in a hierarchical manner, therefore each data set in the hierarchy was considered independently. The experiments carried out with respect to each of the data sets, and the results obtained, are presented in the following subsections.

6.3.7.1 Reuters-21578-COM-2327-1L

The classification results for the Reuters-21578-COM-2327-1L data set are shown in Table 6.15. Reuters-21578-COM-2327-1L comprised 2,327 documents unevenly distributed among 5 classes. The best accuracy result (91.25%) was obtained when SARSET was applied to the first half of the data set, while the best AUC result (0.88) was obtained when SARSET was applied to the entire data set and to the first half of the data set. As in the case of other data sets which contained a small number of classes, good results were obtained in terms of both accuracy and AUC.

Table 6.15: Classification results for the Reuters-21578-COM-2327-1L data set.

	Acc (%)	AUC	Pr	Sn/Re	Sp
Entire data set	91.14	0.88	0.86	0.96	0.87
First half	91.25	0.88	0.86	0.96	0.87
Second half	91.03	0.87	0.86	0.96	0.87

6.3.7.2 Reuters-21578-COM-2327-2L

The classification results for the Reuters-21578-COM-2327-2L data set are shown in Table 6.16. Reuters-21578-COM-2327-2L contained 2,327 documents unevenly distributed among 52 classes. The best accuracy result (85.85%) was obtained when SARSET was applied to the first half of the data set and the best AUC result (0.71) when SARSET

was applied to the entire data set. Note that relatively good results were obtained despite the presence of a large number of classes.

Table 6.16: Classification results for the Reuters-21578-COM-2327-2L data set.

	Acc (%)	AUC	Pr	Sn/Re	Sp
Entire data set	85.56	0.71	0.72	0.99	0.78
First half	85.85	0.70	0.72	0.99	0.78
Second half	85.27	0.70	0.71	0.99	0.77

6.4 Discussion

This section presents a discussion regarding the obtained results and how well they performed with respect to the objectives presented at the beginning of this chapter. Table 6.17 presents an overview of the best results obtained, per data set, using SARSET.

Table 6.17: Best classification results.

Data set	SARSET			CGUSD			Standard Classification		
	Part	Acc (%)	AUC	Best algorithm	Acc (%)	AUC	Best algorithm	Acc (%)	AUC
SAVSNET-840-4-FT	Second half	88.86	0.82	C4.5	55.12	0.75	SMO	89.29	0.95
SAVSNET-971-3-FT	First half	76.13	0.82	RIPPER	62.62	0.60	SMO	74.87	0.80
SAVSNET-917-1L	Second half	74.36	0.78	RIPPER	61.61	0.61	SMO	70.34	0.75
SAVSNET-917-2L	First half	75.42	0.59	*	*	*	RIPPER	66.96	0.57
SAVSNET-917-3L	Second half	71.66	0.74	*	*	*	C4.5	57.25	0.59
SAVSNET-917-4L	Entire data set	52.97	0.07	*	*	*	SMO	47.00	0.64
OHSUMED-CA-3187-1L	Second half	77.21	0.87	SMO	89.10	0.90	SMO	94.84	0.93
OHSUMED-CA-2570-2L	First half	67.78	0.35	RIPPER	73.91	0.90	SMO	80.82	0.94
OHSUMED-CA-834-3L	Second half	82.19	0.61	C4.5	66.88	0.87	SMO	78.42	0.90
OHSUMED-AD-3393-1L	Second half	84.55	0.65	SMO	49.02	0.87	SMO	82.46	0.85
OHSUMED-AD-569-2L	First half	78.84	0.57	SMO	63.33	0.89	C4.5	76.74	0.87
OHSUMED-AD-292-3L	Second half	83.32	0.75	C4.5	85.57	0.96	C4.5	89.69	0.91
Reuters-21578-LOC-2327-1L	First half	79.88	0.59	SMO	41.30	0.84	SMO	82.25	0.93
Reuters-21578-LOC-2327-2L	First half	77.24	0.52	SMO	33.91	0.88	SMO	74.73	0.88
Reuters-21578-COM-2327-1L	First half	91.25	0.88	RIPPER	74.90	0.91	SMO	95.79	0.98
Reuters-21578-COM-2327-2L	First half	85.85	0.70	*	*	*	SMO	84.40	0.97

Prior to considering the objectives of the work described in this chapter the following five observations can be noted:

1. Although no domain experts gave assistance in the experiments carried out, the author's investigations with respect to the various application domains covered by the data sets provided as much expertise as possible to identify the relevant phrases and to use the SARSET semi-automated tool to generate rule based classifiers. In general it can be seen that the results are satisfactory considering the lack of suitable domain experts in each case. It is suggested that the satisfactory results obtained could have been improved further if experienced domain experts had been used. The importance of the involvement of domain experts with respect to the technique presented in this chapter is mainly related to: (i) the identification of more relevant phrases than a user with little, or no knowledge at all, with respect to the domain of interest could identify; and (ii) aiding in the disambiguation of concepts that may not be easily distinguishable by a non-expert user. Consequently, summaries of a better quality may have been generated.
2. SARSET requires the intervention of a domain expert (user) and consequently is more resource intensive. However, given that the author was able to generate sixteen different classifiers using SARSET in reasonable time the resource required does not appear to be a limiting factor. It is difficult to quantify the resource required, as this is dependent on the size (in terms of number of records and classes) and the complexity of the data set to be processed, but each classifier required some 2 hours to generate.
3. Note that there were cases where there was not enough information in the documents regarding certain classes, which can be attributed to the dependency of the different levels in the hierarchical data sets, therefore contributing to the reduction of potential rules.
4. The unbalanced distribution of documents with respect to the classes had repercussions with respect to the performance of SARSET, especially in cases where there was a large number of documents.
5. With regard to the previous observation, many data sets contained a large number of classes. In general, for data sets that had a large number of classes the classification results were not satisfactory; on the other hand for data sets which consisted of a small number of classes the classification results were satisfactory.

With respect to the objectives identified in the introduction to this chapter:

1. *To determine whether the quality of the desired text summarisation classifiers can be improved by incorporating input from domain experts using a semi-automated tool.* From the results obtained it was determined that the quality of the text summarisation classifiers benefit from input by domain experts. As noted above, although in general the results were satisfactory, it is acknowledged that if domain experts had assisted in the generation of the classifiers, much better results might have been obtained.
2. *To ascertain the applicability and effectiveness of the approach when applied to free text from different sources than questionnaires.* The applicability and effectiveness of SARSET when applied to free text from different sources than questionnaires was demonstrated. As already noted, the other sources of text were medical abstracts and news stories. Recall that the type of text on which this thesis is focused is free text from questionnaires and that the approaches presented take this into account.
3. *To compare the operation of the approach proposed in this chapter with those of the previous chapters.* From Table 6.17 it can be seen that SARSET performed well in comparison to the application of standard classification techniques as presented in Chapter 4 and in comparison with CGUSD (Classifier Generation Using Secondary Data), as presented in Chapter 5. Overall, in terms of accuracy, SARSET performed better than when using standard classification techniques and using CGUSD. However, in terms of AUC, SARSET performed worse than using standard classification techniques but as good as CGUSD. SARSET proved to be useful when applied to different sources of text but especially in the case of free text from questionnaires; which, as has been explained before in this thesis, have unique characteristics (unstructured, misspelled words, acronyms, abbreviations) that make it difficult to extract information from them using conventional techniques.

6.5 Summary

This chapter presented SARSET (the Semi-Automated Rule Summarisation Extraction Tool) which supports a semi-automated approach to the generation of text summarisation classifiers. SARSET was evaluated using: (i) the free text element of three questionnaire data sets, (ii) the free text of two data sets containing medical abstracts and (iii) the free text of news stories from a news agency. As in the previous chapter, the evaluation metrics used were: (i) overall accuracy expressed as a percentage, (ii) Area Under the ROC Curve (AUC), (iii) precision, (iv) sensitivity/recall and (v) specificity.

Overall, good classification results were obtained for most of the data sets. In terms of accuracy SARSET obtained better results than when using standard classification techniques and CGUSD. However, in terms of AUC SARSET performed worse than when using standard classification techniques, but was competitive with CGUSD. Recall that most of the data sets were unbalanced and that the identification and selection of relevant phrases from the free text was not made by a domain expert. Regarding this latter consideration it is suggested that the satisfactory results obtained could have been improved if experienced domain experts had generated more comprehensive classifiers. It was thus concluded that SARSET is a technique that has general applicability and can be used in different application domains for the purpose of text summarisation.

Chapter 7

Text Summarisation Using Hierarchical Text Classification

7.1 Introduction

The previous chapter presented a semi-automated classification technique called SARSET (Semi-Automated Rule Summarisation Extraction Tool) which was aimed at conducting document summarisation classification by providing a mechanism for intervention by a domain expert. This chapter presents a hierarchical classification approach for the generation of text summarisation classifiers. The motivation for the approach was that the summary generation processes described in the foregoing three chapters required a collection of classifiers (one per class) so as to generate a comprehensive summary made up of several class labels. Each classifier operates independently of one another. The hierarchical approach presented in this chapter is founded on the observation that the class labels used to produce summarisations are frequently related and hence can be arranged in a *class hierarchy*. The conjecture is that more sophisticated summaries can be produced using the class hierarchy concept so that several class labels can be collectively used to generate the desired summarisations. The hierarchical summarisation classification approach presented in this chapter offers the advantage that different levels of classification can be used and the summarisation customised according to which branch of the tree the current document or questionnaire is located. The approach presented is based on the concept of hierarchical text classification, which is a form of text classification that involves the use of class labels arranged in a tree structure. Hierarchical text classification is thus a form of multi-label classification. As Sun and Lim state [99], hierarchical classification allows a large classification problem to be addressed using a “divide-and-conquer” approach. Hierarchical text classification has been widely investigated and used as an alternative to standard text classification, also known as *flat classification*, where class labels are considered independently from one another.

Recall that, as was described in earlier chapters, the reason why text summarisation

can be conceived of as a form of text classification is that the classes assigned to text documents can be viewed as indications (summarisations) of the main ideas of the text. It has been acknowledged that a summary of this form is not as complete or as extensive as what some practitioners might consider to be a summary; however, as also already noted, if we assign multiple labels to each document, this comes nearer to what might be traditionally viewed as a summary. Note that for a summary to be both informative and complete the number of required classes may be substantial, to the extent that the use of flat classification techniques may no longer be viable. A hierarchical approach to text summarisation classification is therefore suggested. The idea is that by arranging the potential class labels into a hierarchy multiple class labels can be attached to documents in a more effective way than if flat classifiers were used. The effect is to permit an increase in the number of classes that can be used in the summarisation. To the best knowledge of the author hierarchical classification has not previously been studied within the context of text summarisation apart from the author's own work in [30] and [31].

The proposed hierarchical text classification for text summarisation approach offers the following advantages:

1. Humans are used to the concept of defining things in a hierarchical manner, thus the summaries will be produced in an intuitive manner.
2. Hierarchies are a good way of encapsulating knowledge, in the sense that each node that represents a class in the hierarchy has a specific meaning or significance associated with it with respect to the summarisation task.
3. Classification/summarisation can be achieved efficiently without having to consider all class labels for each unseen record.
4. It results in a more effective form of classification/summarisation because it supports the incorporation of specialised classifiers, at specific nodes in the hierarchy.

The five hierarchical data sets used to evaluate the proposed hierarchical summarisation classification technique were:

- SAVSNET-917-4H
 - SAVSNET-917-1L
 - SAVSNET-917-2L
 - SAVSNET-917-3L
 - SAVSNET-917-4L
- OHSUMED-CA-3187-3H

- OHSUMED-CA-3187-1L
- OHSUMED-CA-2570-2L
- OHSUMED-CA-834-3L
- OHSUMED-AD-3393-3H
 - OHSUMED-AD-3393-1L
 - OHSUMED-AD-569-2L
 - OHSUMED-AD-292-3L
- Reuters-21578-LOC-2327-2H
 - Reuters-21578-LOC-2327-1L
 - Reuters-21578-LOC-2327-2L
- Reuters-21578-COM-2327-2H
 - Reuters-21578-COM-2327-1L
 - Reuters-21578-COM-2327-2L

Note that in the above list the different levels of each hierarchy are also listed. Thus, with respect to the original 18 evaluation data sets that were presented in Chapter 3 four of these data sets were not used (SAVSNET-840-4-FT, SAVSNET-840-4-TD+FT, SAVSNET-971-3-FT, SAVSNET-971-3-TD+FT); in other words 14 of the original 18 evaluation data sets are considered in this chapter. The maximum number of levels in the hierarchical data sets used was four, the main reason for this was that it was difficult to obtain hierarchical data sets, suitable for evaluation purposes, with more than four levels. However, the approach presented in this chapter is applicable to hierarchical data sets with higher number of levels. The minimum number of levels for the hierarchical data sets used was two.

The research described in this chapter had four objectives:

1. To determine whether the quality of the desired text summarisation classifiers can benefit from a hierarchical text classification approach.
2. To ascertain the applicability and effectiveness of the approach when applied to free text from different sources than questionnaires.
3. To compare the performance of the summarisation classifiers having hierarchies of classes defined using different criteria (sequential questions, categories of medical topics and location and topics of news stories).

4. To compare the operation of the approach proposed in this chapter with those of the previous chapters and to report any improvements with respect to the classification results obtained in Chapters 4, 5 and 6.

The rest of this chapter is arranged as follows. The hierarchical classification methodology is described in Section 7.2. A description of the hierarchical data sets used is presented in Section 7.3. Section 7.4 presents a comprehensive description of the experiments carried out with respect to each of the hierarchical data sets considered for evaluation purposes. A discussion of how the proposed technique performed is presented in Section 7.5. Finally, a summary of the chapter is presented in Section 7.6.

7.2 Methodology

The proposed methodology is presented in this section, which is divided into three subsections: Subsection 7.2.1 presents a formal definition of the problem that the proposed hierarchical classification technique addresses, Subsection 7.2.2 describes the implementation of the proposed technique in detail, and Subsection 7.2.3 describes how the text summaries are generated using the output of the hierarchical classification.

7.2.1 Problem definition

The input to the proposed text summarisation hierarchical classifier generator is a “training set” of n free text documents, $D = \{d_1, d_2, \dots, d_n\}$, where each document d_i has a sequence of m “summarisation” class labels, $S = \{s_1, s_2, \dots, s_m\}$, such that there is a one-to-one correspondence between each summarisation label s_i and some class g_j . Thus the summarisation labels are drawn from a set of m classes $G = \{g_1, g_2, \dots, g_m\}$ where each class g_j in G has a set of k summarisation labels associated with it $g_i = \{c_{j_1}, c_{j_2}, \dots, c_{j_k}\}$. The associated summary classification hierarchy H then comprises a set of nodes arranged into p levels, $L = \{l_1, l_2, \dots, l_p\}$, such as that shown in Figure 7.1. Except at the leaf nodes, each node in the hierarchy has a classifier associated with it. The leaf nodes hold the classes that do not have associated subclasses, therefore they are not used for classifier generation.

Since we are using a top-down model, the classifiers can be arranged according to two approaches in terms of the scope and dependency between levels: (i) cascading and (ii) non-cascading. In the cascading case, the output of the classifier at a parent node influences the classification conducted at the child nodes at the next level of the hierarchy (we say that the classification process “cascades” downwards). Thus a classifier is generated for each child node (except the leaf nodes) depending on the resulting classification from the parent node. The classification process continues in this manner until there are no more nodes to be developed. In the case of the non-cascading model each classifier is generated independently from that of the parent node.

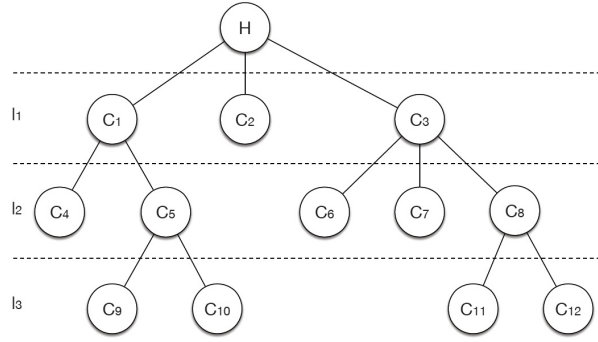


Figure 7.1: Hierarchy of classes.

In addition, two types of hierarchies are identified regarding the parent-child node relationship: single and multi-parent. The top-down strategy can be applied in both cases because, given a piece of text to be summarised, only one best child node (class) is selected per level. Examples of single and multi-parent hierarchies are shown in Figures 7.2 and 7.3.

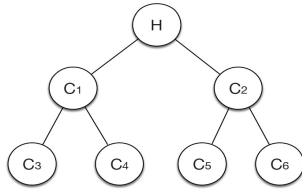


Figure 7.2: Single-parent hierarchy

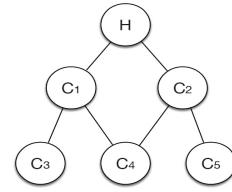


Figure 7.3: Multi-parent hierarchy

The classifier generation process is closely linked to the structure of the hierarchy. When generating a cascading hierarchy we start by generating a classifier founded on the entire training set. For its immediate child branches we only used that part of the training set associated with the class represented by each child node. For non-cascading we use the entire training set for all nodes (except the leaf nodes), but with different labels associated with the training documents according to the level we are at. The classifier generation process is described in more detail in Subsection 7.2.2.

Once the generation process is complete, the text summarisation classifier is ready for application to new data. New documents will be classified by traversing the hierarchical classifier in a similar manner to that used in the context of a decision tree. That is, at each level the process will be directed towards a particular branch in the hierarchy according to the current classification. The length of the produced summary will depend on the number of levels traversed within the hierarchy.

7.2.2 Hierarchical Classification

In this subsection the two hierarchical text classifier generation approaches considered (cascading and non-cascading) are described in more detail. Recall that in the non-

cascading approach the classification process is carried out independently in each node and, as its name implies, independently of the levels and the parent-child node relationship, in other words, flat classifiers are generated for each node; in the cascading approach the output of the classification of the parent nodes affects the classification conducted at child nodes at the next level of the hierarchy. In both cases a five-step classifier generation process is specified, as follows:

1. **Preprocessing of documents:** Text is converted to lower case; numbers, symbols and stop words (common words that are not significant for the text classification/summarisation process) are removed, stemming is applied and feature selection is performed.
2. **Classification of documents:** A classifier is generated for each node in the current level of the hierarchy. The nature of the classification depends on the approach taken:
 - (i) **Cascading approach:** The output of the classification of the nodes in the current level will affect nodes at the next level down in the hierarchy.
 - (ii) **Non-cascading approach:** The classifier generation is carried out independently in each node. The classification results do not affect the classifier generation in nodes at the next level down in the hierarchy.
3. **Evaluation of the classification:** The generated classifier is evaluated (so the confidence in the classifier can be gained) and the results recorded. Based on the resulting evaluation metrics:
 - (i) **Cascading approach:** Correctly and incorrectly classified instances are considered for the classification process in the next level down in the hierarchy.
 - (ii) **Non-cascading approach:** The classification results from the previous level are not taken into account for the classifier generation conducted at the next level down in the hierarchy.
4. **Verification of the existence of nodes in the next level down in the hierarchy:** Exit (hierarchical classifier is complete) if there are no more nodes to be developed at the next level down in the hierarchy. Otherwise continue with step 5.
5. **Classifier generation at the next level down in the hierarchy:** Repeat process from step 2 for each node at the next level down in the hierarchy.
 - (i) **Cascading approach:** Correctly and incorrectly classified instances for nodes in the previous level are taken into account for nodes in the current level.

(ii) **Non-cascading approach:** Classifier generation at nodes in the current level will be performed regardless of the classification results produced at the previous level.

Once complete (and found to be effective when applied to an appropriately defined test set) the generated classifier may be applied to unseen data.

7.2.3 Summary Generation

The hierarchical classifier, produced as described above, may then be used for summary generation. Similarly to the approaches presented in Chapters 4, 5 and 6 rules defined by domain experts are used to prepend or append domain-specific text to assigned class labels resulting from the classification process. Note, however, that in the hierarchical case more than one class label is always used to generate a summary. An example of such a summarisation, with respect to a hierarchy of four levels, where after the hierarchical classification four summarisation labels ($g_i = \{c_{j_1}, c_{j_2}, c_{j_3}, c_{j_4}\}$) would be related to a document (one per level), might be: *{ “This document was about c_{j_1} with c_{j_2} . c_{j_3} was presented with c_{j_4} ”}*. Note that such a summarisation differs from traditional text summarisation techniques in that the words or phrases that comprise the resulting summary are not necessarily present in the original text. However, the resulting summary is considered to be a concise, coherent and informative overview of the content of a document as confirmed by reference to domain experts in the context of the SAVSNET data.

7.3 Hierarchical data sets used

This section presents a description of the hierarchical data sets used. Note that although a general description of all the data sets was presented in Chapter 3 a more detailed description is presented in this section in order to put all the data sets in the context of their respective hierarchies.

7.3.1 SAVSNET-917-4H

The SAVSNET-917-4H hierarchy consists of 917 documents and several class attributes arranged over 4 different levels in a class hierarchy defined by specific questions variously included in the SAVSNET questionnaires. Note that since the questions were asked sequentially, there is a dependency from higher to lower levels in the hierarchy of classes in terms of the specific characteristics of the conditions presented. Recall that the SAVSNET initiative relates to the frequency of the occurrence of small animal diseases. Although SAVSNET-917-4H represents a relatively small hierarchy of classes, it is the hierarchy of classes with the most levels with respect to the research described in this thesis. The data set is concerned with gastrointestinal symptoms. In this

particular case, the child nodes for each of the gastrointestinal symptoms are similar in the sense that they can have any parent of the previous level (one parent per child node), making the hierarchy symmetric. This hierarchical data set therefore comprised four sub-data sets associated with each level in the hierarchy:

1. **SAVSNET-917-1L**: Contains 3 classes and 917 documents. This level of the hierarchy is related to gastrointestinal symptoms. The distribution of documents per class is presented in Table A.3 of Appendix A.
2. **SAVSNET-917-2L**: Contains 3 classes and 917 documents. This level of the hierarchy is related to the severity of the gastrointestinal symptoms of SAVSNET-917-1L in terms of the presence of haemorrhages. The distribution of documents per class is presented in Table A.4 of Appendix A.
3. **SAVSNET-917-3L**: Contains 3 classes and 917 documents. This level of the hierarchy is related to the occurrence of the gastrointestinal symptoms of the first level taking into account the severity of the second level. The distribution of documents per class is presented in Table A.5 of Appendix A.
4. **SAVSNET-917-4L**: Contains 5 classes and 917 documents. This level of the hierarchy is related to the duration of the gastrointestinal symptoms of the first level. The distribution of documents per class is presented in Table A.6 of Appendix A.

7.3.2 OHSUMED-CA-3187-3H

The OHSUMED-CA-3187-3H hierarchy consists of 3,187 documents and several class attributes arranged over 3 different levels in a class hierarchy related to Cardiovascular Abnormalities found in the MEDLINE database as defined by MeSH tree codes. This hierarchy of classes goes from general diseases at the higher levels to more specific diseases at the lower levels. In this case not all the parent nodes have children making the hierarchy asymmetric. This hierarchical data set contained three sub-data sets associated with each level in the hierarchy:

1. **OHSUMED-CA-3187-1L**: Contains 2 classes and 3,187 documents. This level of the hierarchy is related to the main cardiovascular abnormalities. The distribution of documents per class is presented in Table A.7 of Appendix A.
2. **OHSUMED-CA-2570-2L**: Contains 16 classes and 2,570 documents. This level of the hierarchy is related to more specific cardiovascular abnormalities. The distribution of documents per class is presented in Table A.8 of Appendix A.

3. **OHSUMED-CA-834-3L**: Contains 7 classes and 834 documents. This level of the hierarchy is related to even more specific cardiovascular abnormalities. The distribution of documents per class is presented in Table A.9 of Appendix A.

7.3.3 OHSUMED-AD-3393-3H

The OHSUMED-AD-3393-3H hierarchy consists of 3,393 documents and several class attributes arranged over 3 different levels in a class hierarchy related to Animal Diseases found in the MEDLINE database as defined by MeSH tree codes. Similarly to the OHSUMED-CA-3187-3H hierarchy, this hierarchy of classes goes from general diseases at the higher levels to more specific diseases at the lower levels. Also not all the parent nodes in the OHSUMED-AD-3393-3H hierarchy have children, making the hierarchy asymmetric. This hierarchical data set comprised three sub-data sets associated with each level in the hierarchy:

1. **OHSUMED-AD-3393-1L**: Contains 34 classes and 3,393 documents. This level of the hierarchy is related to animal diseases in general. The distribution of documents per class is presented in Table A.10 of Appendix A.
2. **OHSUMED-AD-569-2L**: Contains 26 classes and 569 documents. This level of the hierarchy is related to more specialised animal diseases. The distribution of documents per class is presented in Table A.11 of Appendix A.
3. **OHSUMED-AD-292-3L**: Contains 9 classes and 292 documents. This level of the hierarchy is related to even more specialised animal diseases. The distribution of documents per class is presented in Table A.12 of Appendix A.

7.3.4 Reuters-21578

The Reuters-21578 collection used in this thesis consists of 2,327 documents and several class attributes arranged over two different levels of two unrelated class hierarchies, but related to the same news stories from the Reuters news agency. As mentioned in Chapter 3, the hierarchy of classes was defined according to the class labels associated with the documents. These class labels were not explicitly organised as a hierarchy, thus an appropriate hierarchy had to be imposed. Two different hierarchies of classes were identified from the class labels used in the Reuters-21578 data set: (i) Reuters-21578-LOC-2327-2H and (ii) Reuters-21578-COM-2327-2H. The first was related to the location where the news stories were reported, the second was related to the topics on which the news stories report, more specifically types of commodities used in trade. Both the Reuters-21578-LOC-2327-2H and Reuters-21578-COM-2327-2H hierarchies are asymmetric hierarchies. Note that both two-level hierarchies refer to the same documents, so the summaries will consist of four class labels (two per class hierarchy). In the following subsections the Reuters-21578 hierarchies are described in more detail.

7.3.4.1 Reuters-21578-LOC-2327-2H

The Reuters-21578-LOC-2327-2H data consists of 2,327 documents and several class attributes arranged over two levels in a class hierarchy related to the location where the news stories were reported:

1. **Reuters-21578-LOC-2327-1L:** Contains 14 classes and 2,327 documents. This level of the hierarchy is related to the region(s) or continent(s) with which each news story was related. In this case there were no class labels that explicitly indicated the region(s) or continent(s); the name of the associated region(s) or continent(s) was determined by taking into account the country or countries with which each news story was related. The distribution of documents per class is presented in Table A.13 of Appendix A.
2. **Reuters-21578-LOC-2327-2L:** Contains 92 classes and 2,327 documents. This level of the hierarchy is related to the country or countries with which each news story was related. The class labels explicitly indicated the country or countries. The distribution of documents per class is presented in Table A.15 of Appendix A.

7.3.4.2 Reuters-21578-COM-2327-2H

The Reuters-21578-COM-2327-2H data consists of 2,327 documents and several class attributes arranged over two levels in a class hierarchy related to the topics to which the news stories reported on:

1. **Reuters-21578-COM-2327-1L:** Contains 5 classes and 2,327 documents. This level of the hierarchy is related to the types of commodities to which the news stories report on. Similarly to the first level of the Reuters-21578-LOC-2327-2H hierarchy, the names of the types of commodities were not explicitly indicated in the documents, therefore the names of the individual commodities, which were explicitly indicated in the documents, were used to determine the types of commodities. The distribution of documents per class is presented in Table A.14 of Appendix A.
2. **Reuters-21578-COM-2327-2L:** Contains 52 classes and 2,327 documents. This level of the hierarchy is related to the individual commodities to which the news stories report on. The class labels associated with the individual commodities in this case were explicitly indicated in the documents. The distribution of documents per class is presented in Table A.16 of Appendix A.

7.4 Experiments and Results

As stated at the beginning of this chapter, the evaluation of the proposed hierarchical approach for generating summaries from free text was carried out using the five hierarchical data sets: (i) SAVSNET-917-4H, (ii) OHSUMED-CA-3187-3H, (iii) OHSUMED-AD-3393-3H, (iv) Reuters-21578-LOC-2327-2H and (v) Reuters-21578-COM-2327-2H, which were described in Section 7.3. Note that in most of the data sets the distribution of the documents per class was significantly unbalanced. The free text was preprocessed by converting it to lower case, removing numbers, symbols and stop words, applying stemming using an implementation of the Porter Stemming algorithm [106] and selecting relevant features using an implementation of the Chi-squared method as described previously in Chapter 3. Similarly to Chapter 5, three classifier generation mechanisms were considered, the ones that produced the best performance as established by the experiments reported in Chapter 4, namely: (i) SMO, (ii) C4.5 and (iii) RIPPER.

As stated in the survey of hierarchical classification presented by Silla and Freitas in [95], there is still not a general consensus on which is the best way to evaluate hierarchical classification algorithms. Most researchers use standard flat classification evaluation measures (accuracy, AUC, precision, sensitivity/recall and specificity for example) to evaluate hierarchical classification, which are the most used with respect to other types of classification but which might not give enough insight into the results for each level in a hierarchy of classes. However, other researchers have presented their own hierarchical classification evaluation measures whose usage is not as widespread. Having taken into account the aforementioned points and considering that the approaches presented in Chapters 4, 5 and 6 were evaluated using flat classification evaluation measures, it was decided to use the same flat classification evaluation measures used in previous chapters to evaluate the hierarchical classification approach used for generating summaries from free text as presented in this chapter. The evaluation was thus conducted using Ten-fold Cross Validation (TCV) and, as in the case of previous chapters, the evaluation metrics used were: (i) overall accuracy expressed as a percentage, (ii) Area Under the ROC Curve (AUC), (iii) precision, (iv) sensitivity/recall and (v) specificity. Again, accuracy and AUC were considered to be the most relevant evaluation measures; precision, sensitivity/recall and specificity were recorded so as to provide a broader insight into the effectiveness of the individual classifiers. (Each of these measures was described in Chapter 2.) Note that the results that are presented are from the parent nodes at each respective level in each one of the hierarchy of classes considered.

Given that many parameters were taken into account for the experiments, 15 tables were produced. In this section only two of these 15 tables are presented to show the best results for the SAVSNET-917-4H (see Table 7.1) and the OHSUMED-CA-3187-3H (see Table 7.2) hierarchies, the former related to free text from questionnaires and the latter to free text from medical abstracts. The complete set of 15 results tables are

presented in Appendix B from Tables B.1 to B.15. For all the tables the first row indicates which classification algorithm was used (SMO, C4.5 or RIPPER), the second row indicates the hierarchical classification approach that was used (cascading or non-cascading) and the third row is the header of the results table. With respect to the latter row, the first column presents the level of the hierarchy and the second column presents the nodes from which the results were obtained. The remaining ten columns are divided in two groups of five columns according to the hierarchical classification strategy applied (cascading and non-cascading respectively); for each group the results obtained are presented in terms of the evaluation measures used: (i) overall accuracy expressed as a percentage (Acc), (ii) Area Under the ROC Curve (AUC), (iii) precision (Pr), (iv) sensitivity/recall (Sn/Re) and (v) specificity (Sp).

Table 7.1: Classification results for SAVSNET-917-4H using SMO.

		SMO									
Algorithm		casc					¬casc				
Approach		Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
Level	Node										
First	GI Diseases	70.34	0.75	0.68	0.70	0.75	70.34	0.75	0.68	0.70	0.75
Second	Diarrhoea	64.01	0.62	0.61	0.64	0.59	74.63	0.68	0.74	0.75	0.61
	Vomiting	75.43	0.52	0.68	0.75	0.29	89.11	0.49	0.82	0.89	0.10
	Vomiting & Diarrhoea	55.21	0.60	0.56	0.55	0.63	98.50	0.50	0.97	0.99	0.02
Third	Haemorrhagic	69.17	0.64	0.63	0.69	0.58	61.02	0.59	0.57	0.61	0.56
	Not Haemorrhagic	61.10	0.60	0.59	0.61	0.57	63.08	0.59	0.59	0.63	0.53
	Unknown Severity	68.06	0.58	0.64	0.68	0.47	58.82	0.61	0.57	0.59	0.62
Fourth	First Time	51.18	0.62	0.49	0.51	0.66	47.99	0.59	0.44	0.48	0.66
	Nth Time	37.34	0.60	0.36	0.37	0.75	43.10	0.64	0.38	0.43	0.75
	Unknown Occurrence	*	*	*	*	*	51.85	0.46	0.39	0.52	0.40

Table 7.2: Classification results for OHSUMED-CA-3187-3H using SMO.

Algorithm	SMO										
Approach											
Level	Node	casc					¬casc				
		Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
First	Types of CA	94.84	0.93	0.95	0.95	0.91	94.84	0.93	0.95	0.95	0.91
Second	Congenital Heart Defects Vascular Malformations	81.36 89.01	0.93 0.88	0.83 0.89	0.81 0.89	0.94 0.87	79.95 92.58	0.94 0.92	0.81 0.92	0.80 0.93	0.96 0.91
Third	Heart Septal Defects Vascular Fistula	77.81 96.36	0.83 0.56	0.78 0.97	0.78 0.96	0.83 0.15	80.98 94.40	0.88 0.73	0.75 0.94	0.81 0.94	0.92 0.52

Note that for identifying the hierarchical strategy and the classification algorithm that had the best performance with respect to each level of the hierarchy of classes, and for allowing comparison of the approach presented in this chapter with the approaches presented previously in this thesis, the results obtained had to be presented in the closest manner possible to those presented in Chapters 4, 5 and 6. It was thus decided to average the accuracy and AUC results obtained per level for each hierarchy of classes in a similar manner as other researchers have presented averaged values for evaluating hierarchical classification ([20, 53, 66, 99]). Simplified results tables are presented in Tables 7.3 to 7.6 with respect to each hierarchy of classes except for OHSUMED-AD-3393-3H, which turned out to be not completely suitable for the experiments given its unbalanced nature and its distribution of documents per parent node, thus not allowing the comparison of the cascading and non-cascading hierarchical strategies. The simplified results tables are organised in the following manner: the first column indicates the level of the hierarchy, the remaining 12 columns are divided into groups of 4 columns (one per classification algorithm) which in turn are divided into groups of 2 columns (one per hierarchical classification strategy) which in each case give the averaged accuracy and averaged AUC results per level.

Table 7.3: Averaged accuracy and AUC results for SAVSNET-917-4H using SMO, C4.5 and RIPPER.

Level	SMO				C4.5				RIPPER			
	casc		¬casc		casc		¬casc		casc		¬casc	
	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC
First	70.34	0.75	70.34	0.75	63.14	0.69	63.14	0.69	67.18	0.69	67.18	0.69
Second	64.88	0.58	87.41	0.56	56.48	0.56	85.50	0.41	53.81	0.53	88.43	0.41
Third	66.11	0.60	60.97	0.59	56.66	0.55	50.46	0.51	60.73	0.51	59.29	0.53
Fourth	44.26	0.61	45.55	0.61	36.72	0.54	34.81	0.54	41.63	0.51	43.38	0.54

Table 7.4: Averaged accuracy and AUC results for OHSUMED-CA-3187-3H using SMO, C4.5 and RIPPER.

Level	SMO				C4.5				RIPPER			
	casc		¬casc		casc		¬casc		casc		¬casc	
	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC
First	94.84	0.93	94.84	0.93	93.16	0.91	93.16	0.91	93.38	0.90	93.38	0.90
Second	85.19	0.90	86.27	0.93	84.90	0.89	84.45	0.91	87.53	0.91	83.89	0.90
Third	87.09	0.69	87.69	0.81	85.95	0.73	82.58	0.79	89.08	0.67	84.12	0.79

Table 7.5: Averaged accuracy and AUC results for Reuters-21578-LOC-2327-2H using SMO, C4.5 and RIPPER.

Level	SMO				C4.5				RIPPER			
	casc		¬casc		casc		¬casc		casc		¬casc	
	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC
First	82.25	0.93	82.25	0.93	73.79	0.83	73.79	0.83	71.85	0.84	71.85	0.84
Second	76.68	0.69	70.49	0.71	60.19	0.71	67.85	0.77	69.07	0.65	72.62	0.74

Table 7.6: Averaged accuracy and AUC results for Reuters-21578-COM-2327-2H using SMO, C4.5 and RIPPER.

Level	SMO				C4.5				RIPPER			
	casc		¬casc		casc		¬casc		casc		¬casc	
	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC
First	95.79	0.98	95.79	0.98	88.27	0.94	88.27	0.94	91.41	0.96	91.41	0.96
Second	81.64	0.83	79.96	0.87	73.31	0.80	81.37	0.87	76.01	0.80	78.83	0.84

The rest of this section is divided into six subsections: the first five directed at each one of the hierarchies of classes considered and the last one directed at the resulting text summaries.

7.4.1 SAVSNET-917-4H

The hierarchical classification results for the SAVSNET-917-4H hierarchy of classes are presented in detail in Appendix B in Tables B.1, B.2 and B.3, and in a simplified form in Table 7.3. Note that previous work by the author regarding hierarchical classification for text summarisation that included a SAVSNET hierarchy of classes was presented in [30]; the main differences with respect to that work are: (i) slightly different pre-processing, (ii) the addition of C4.5 and RIPPER as classification algorithms and (iii) the addition of precision as an evaluation measure. As has been mentioned in previous chapters, the documents are unevenly distributed among the different classes in the hierarchy. Recall that the SAVSNET-917-4H hierarchy of classes was defined in terms of sequential questions from the SAVSNET project, and consequently all the nodes, except for the ones at the lowest level, were considered. For the three classification algorithms and for both the cascading and non-cascading strategies the classification results for both accuracy and AUC were better at higher levels and worst at lower levels. Table 7.1 presents the detailed results obtained when SMO was used, which produced the best results, but still shows how both accuracy and AUC decreased from the higher to the lower levels in the hierarchies. Note that with respect to the cascading strategy one node at the lowest level in the hierarchy (*“Unknown Occurrence”*) did not get documents from higher levels in the hierarchy, which was caused by the misclassification of a small number of records related to that class. In general the results were not satisfactory, with SMO obtaining the best results for both the cascading and the non-cascading strategies.

7.4.2 OHSUMED-CA-3187-3H

The hierarchical classification results for the OHSUMED-CA-3187-3H hierarchy of classes are presented in detail in Appendix B in Tables B.4, B.5 and B.6, and in a simplified form in Table 7.4. Recall that since the OHSUMED-CA-3187-3H hierarchy of classes was defined according to the categories in MEDLINE’s MeSH tree only the parent nodes (classes) in each level of the hierarchy were considered for classifier generation, thus, reducing the number of classes to be used. Similarly to the results obtained for the SAVSNET-917-4H hierarchy, better results were obtained at the higher levels of the hierarchy, decreasing lower down in the hierarchy. However, as Table 7.4 shows, the results were satisfactory in this case. Overall the best results were obtained using SMO and the non-cascading hierarchical strategy (see Table 7.2). In terms of accuracy the results obtained with both the cascading and non-cascading strategies were similar.

In terms of AUC the best results were obtained with the non-cascading strategy.

7.4.3 OHSUMED-AD-3393-3H

The hierarchical classification results for the OHSUMED-AD-3393-3H hierarchy of classes are presented in detail in Appendix B in Tables B.7, B.8 and B.9. The results were not presented in a simplified form here because, as indicated by closer inspection of the results, the OHSUMED-AD-3393-3H hierarchy of classes turned out to be not entirely suited to the experiments because of the unbalanced nature of the data set and the small number of documents in the parent nodes; which, unlike the other hierarchical data sets used, did not allow for passing (cascading) of sufficient numbers of documents to lower levels in the hierarchy. In the case of the non-cascading strategy results were obtained for all the parent nodes in the hierarchy with the exception of three nodes in the second level of the hierarchy, where there were not enough documents to perform TCV. In terms of the non-cascading strategy the classification algorithm that had the best performance was SMO. Consequently, since few results were obtained using the cascading strategy for the three classification algorithms considered, it was not possible to make a comparison between the cascading and non-cascading strategies.

7.4.4 Reuters-21578-LOC-2327-2H

The hierarchical classification results for the Reuters-21578-LOC-2327-2H hierarchy of classes are presented in detail in Appendix B in Tables B.10, B.11 and B.12, and in a simplified form in Table 7.5. Note that it was not possible to obtain results for some nodes of this two-level hierarchy of classes because of the small number of documents related to that respective nodes (classes) and also because in some cases there were not enough documents to perform TCV. Overall the best results for all the classification algorithms were obtained using the non-cascading strategy. In terms of both accuracy and AUC the best results were obtained using SMO.

7.4.5 Reuters-21578-COM-2327-2H

The hierarchical classification results for the Reuters-21578-COM-2327-2H hierarchy of classes are presented in detail in Appendix B in Tables B.13, B.14 and B.15, and in a simplified form in Table 7.6. Given the slightly more even distribution of documents among a small number of classes it was possible to obtain results with respect to all the nodes in this two-level hierarchy of classes. In general the accuracy and AUC results were relatively good with respect to the hierarchical classification strategies and to the classification algorithms. The best results were obtained using SMO and the cascading strategy.

7.4.6 Text Summarisation

In this subsection some examples of generated summaries for each data collection considered (SAVSNET, OHSUMED and Reuters-21578) are presented. Recall that, as noted in Subsection 7.2.3, summaries are generated by prepending or appending domains-specific text to the class labels assigned to the documents after the classification process. Therefore the size of the generated summaries depends on the number of levels considered either using a hierarchy of classes per document set (SAVSNET, OHSUMED) or two or more hierarchies of classes per document set (Reuters-21578). This subsection is divided into three parts, one per data collection considered. It was divided in this manner so as to group data sets containing similar types of free text.

7.4.6.1 SAVSNET

With respect to the SAVSNET data collection a four-level hierarchical data set was considered, namely SAVSNET-917-4H. Summaries for this hierarchical data set in terms of the topic of each level are of the form: $\{GI \text{ (Gastrointestinal) disease presented, severity of the GI disease presented, occurrence of the GI disease presented, duration of the GI disease presented}\}$. Note that, if desired, domain-specific text related to veterinary science can be added based on rules defined by domain experts in order to improve the readability of the generated summary. Consider, for example, the following resulting class labels, each one related to a level in the hierarchy: $\{diarrhoea, not haemorrhagic, first time, 2-4\}$. A summary based on these class labels could be: $\{Presented diarrhoea, not haemorrhagic, presented for the first time and the duration of the symptoms was between two and four days.\}$. Note that the hierarchy of classes was defined using a sequence of questionnaire questions, therefore in most cases all the levels in the resulting hierarchy are covered. In cases where there are insufficient documents at certain nodes, the length of the summaries will vary, especially in the case where the cascading hierarchical strategy is used because there may not be a sufficient number of documents to pass to the next level.

7.4.6.2 OHSUMED

For the OHSUMED data collection two hierarchical data sets were considered: (i) OHSUMED-CA-3187-3H and (ii) OHSUMED-AD-3393-3H. Although both data sets are subsets of the OHSUMED data collection there is no close relationship between them, as they are part of different branches of the MEDLINE's MeSH tree which reflects how the content of MEDLINE (medical abstracts) is organised. Therefore, they were considered as individual and separate data sets. As noted previously, the OHSUMED-CA-3187-3H hierarchy of classes is related to Cardiovascular Abnormalities and the OHSUMED-AD-3393-3H hierarchy of classes to Animal Diseases. An example of a generated summary from OHSUMED-CA-3187-3H going from the general to

the particular (higher to lower classes) with the class labels $\{Vascular\ Malformations, Vascular\ Fistula, Arteriovenous\ Fistula\}$ would be: $\{Paper\ related\ to\ Vascular\ Malformations, Vascular\ Fistula\ and\ Arteriovenous\ Fistula\}$. In this example the summary is a list of topics to which the documents relate, going from the general to the specific. A domain expert would be able to add more meaningful text to be prepended or appended as appropriate. An example of a generated summary from OHSUMED-AD-3393-3H going from the particular to the general (lower to higher classes) with the class labels $\{Animal\ Parasitic\ Diseases, Animal\ Protozoan\ Infections, Animal\ Toxoplasmosis\}$ would be: $\{Paper\ related\ to\ Animal\ Toxoplasmosis\ and\ Animal\ Protozoan\ Infections\ in\ the\ context\ of\ Animal\ Parasitic\ Diseases\}$. Since the OHSUMED hierarchies of classes are asymmetric and do not necessarily cover all the levels in the hierarchy, the generated summaries are more varied in terms of length.

7.4.6.3 Reuters-21578

The Reuters-21578 data set was distinct from the SAVSNET and OHSUMED data sets because the summaries were generated from two unrelated two-level hierarchies of classes related to the same documents. The two hierarchical data sets considered were: (i) Reuters-21578-LOC-2327-2H and (ii) Reuters-21578-COM-2327-2H. The former is related to the location from where the associated news story was reported and the latter to the topics of the news stories (namely commodities used in trade). Since the same documents were related to both hierarchies of classes, it was deemed appropriate to construct the summaries using the four class labels from both hierarchies (two class labels per hierarchy). This presents a different approach to generating summaries from those used with respect to the SAVSNET and the OHSUMED data collections. An example of a generated summary for a certain document with respect to the Reuters-21578-LOC-2327-2H and Reuters-21578-COM-2327-2H data sets with the class labels $\{Europe, UK\}$ and $\{Energy, Crude\}$ respectively would be: $\{This\ news\ story\ is\ about\ Crude\ and\ Energy, and\ is\ related\ to\ the\ UK\ in\ the\ Europe\ area\}$. Note that this case allows for the combination of the assigned class labels in a number of different ways. It is acknowledged that having hierarchies with more levels would have resulted in more comprehensive summaries.

7.5 Discussion

This section presents a discussion regarding the obtained results and how well they performed with respect to the objectives presented at the beginning of this chapter and with respect to the results obtained in previous chapters. Table 7.7 presents an overview of the best results obtained per level for each hierarchy of classes for which results could be obtained. Note that the classification results for the OHSUMED-AD-3393-3H hierarchy of classes are not included because it was not possible to obtain

results using the cascading strategy (see above). A comparison between the results obtained with the hierarchical classification approach presented in this chapter and the results obtained with approaches presented in Chapters 4, 5 and 6 is presented along with Table 7.8, which presents the overall classification results for the approaches presented in this thesis with respect to the evaluation data sets considered in each case.

Table 7.7: Best classification techniques and results.

Data set	Best results					
	casc			\neg casc		
	Alg.	Acc (%)	AUC	Alg.	Acc (%)	AUC
SAVSNET-917-1L	SMO	70.34	0.75	SMO	70.34	0.75
SAVSNET-917-2L	SMO	64.88	0.58	SMO	87.41	0.56
SAVSNET-917-3L	SMO	66.11	0.60	SMO	60.97	0.59
SAVSNET-917-4L	SMO	44.26	0.61	SMO	45.55	0.61
OHSUMED-CA-3187-1L	SMO	94.84	0.93	SMO	94.84	0.93
OHSUMED-CA-2570-2L	RIPPER	87.53	0.91	SMO	86.27	0.93
OHSUMED-CA-834-3L	C4.5	85.95	0.73	SMO	87.69	0.81
Reuters-21578-LOC-2327-1L	SMO	82.25	0.93	SMO	82.25	0.93
Reuters-21578-LOC-2327-2L	SMO	76.68	0.69	RIPPER	72.62	0.74
Reuters-21578-COM-2327-1L	SMO	95.79	0.98	SMO	95.79	0.98
Reuters-21578-COM-2327-2L	SMO	81.64	0.83	C4.5	81.37	0.87

Prior to considering the objectives of the work described in this chapter the following five observations can be made:

1. For evaluation purposes five hierarchies of classes were used. The hierarchies of classes were subsets of the data collections used earlier in this research, namely: (i) SAVSNET (SAVSNET-917-4H), (ii) OHSUMED (OHSUMED-CA-3187-3H and OHSUMED-AD-3393-3H) and (iii) Reuters-21578 (Reuters-21578-LOC-2327-2H and Reuters-21578-COM-2327-2H). The way in which the hierarchies of classes used were defined was based on the nature of the data collections from which they were part of: (i) the SAVSNET-917-4H hierarchy was defined in terms of the sequential questions itemised in the SAVSNET questionnaires, (ii) the OHSUMED-CA-3187-3H and the OHSUMED-AD-3393-3H hierarchies were defined according to MeSH tree codes from the MEDLINE biomedical database, and (iii) the Reuters-21578-LOC-2327-2H and Reuters-21578-COM-2327-2H hi-

erarchies were defined in terms of the locations and topics (commodities used in trade) to which news stories pertained.

2. Regarding the SAVSNET and OHSUMED hierarchies, only the current single hierarchy was considered for the generation of summaries. However, in the case of the Reuters-21578 hierarchies, since they were related to the same set of documents but were unrelated in terms of how they were defined, the logical step was to apply the hierarchical method for each hierarchy and to use the output of both hierarchies for summary generation. In this particular case, although both Reuters-21578 hierarchies had two levels, it was possible to generate comprehensive summaries when the classification outputs were combined.
3. Most of the data sets associated with the levels in the hierarchies had an uneven distribution of documents per class, which in some cases affected severely the operation of the cascading strategy because: (i) there were not enough documents to be considered in some levels in the hierarchy (usually the lowest ones) and (ii) unbalanced data sets could lead to misclassifications, some of which, in the case of the cascading strategy, were carried forward to the lower levels in the hierarchy.
4. Since there is not an established way to evaluate hierarchical classification systems and in order to be able to compare the results of this chapter to those obtained in Chapters 4, 5 and 6, the same flat classification evaluation measures used in those chapters were used to evaluate the hierarchical classification approach presented in this chapter. Note that the accuracy and AUC results were averaged for each level of the hierarchies as has been done in other related work found in the literature.
5. As expected, the quality of the hierarchical classification results obtained was reflected in the quality of the generated summaries. The summary generation with respect to the hierarchical approach presented in this chapter was carried out by appending or prepending domain-specific text to the resulting class labels (one per level) in a similar manner as in previous chapters. Comprehensive summaries were generated given that many class labels were used, unlike the approaches presented in previous chapters which only used one class label although more comprehensive results could have been obtained by harnessing the output of several classifiers each directed at a different set of class labels. Additionally, the hierarchical approach presented in this chapter allows for the construction of summaries by combining the resulting class labels in the most convenient way by appending/prepending domain-specific text that was previously defined by domain experts to, for example, generate summaries that go from the general to the particular (or the other way around).

The objective of comparing the evaluation results for all the approaches with respect to the data sets used was to identify which approach produced the best performance with respect to the data sets considered. Since the proposed approaches operated in different manners and were, at least in part, intended for use in different contexts, the comparison presented here is not intended to identify one approach as being superior to the rest, instead the intention is to identify the features of each that lead to effective summarisation.

The comparison of the proposed approaches was primarily conducted in terms of classification performance. For all the approaches the evaluation metrics used were: (i) overall accuracy expressed as a percentage, (ii) Area Under the ROC Curve (AUC), (iii) precision, (iv) sensitivity/recall and (v) specificity. Accuracy and AUC were considered to be the most relevant; precision, sensitivity/recall and specificity were presented so as to provide a broader insight into the effectiveness of the approaches. The reason for considering accuracy and AUC as the most relevant evaluation metrics were as follows. Accuracy is the simplest and easiest measure to understand of the evaluation measures used since it is a percentage given by the number of instances correctly classified, however accuracy does not take into account the distribution of the classes (the “class priors”). Thus AUC was also used as this does take into account the class priors. Note that with respect to the evaluation of the hierarchical approach presented in this chapter, the evaluation metrics were calculated for each node at each level of the hierarchy, but for comparison purposes the overall accuracy and AUC results were obtained by averaging for each level.

How the division of the evaluation data into training and test components was conducted depended on the nature of the proposed text summarisation technique under consideration. For the approach using standard classification techniques presented in Chapter 4, Ten-fold Cross Validation (TCV) was used. For the CGUSD approach presented in Chapter 5 each classification technique was trained using secondary data and tested using primary data. For the SARSET approach presented in Chapter 6, because of the human resource required, a 50:50 training-set split was adopted where the classifiers were applied to the test set, the original training sets and the entire data set (training and test set) to ensure that all the documents were used for the evaluation. For the hierarchical classification technique presented in this chapter TCV was again used.

Table 7.8 presents the overall classification results for the approaches presented in this thesis with respect to the evaluation data sets considered in each case. The results are presented in terms of classification accuracy and AUC. The first column of the table gives the name of the data set (individually for standard classification, CGUSD and SARSET, and as levels in a hierarchy of classes with respect to the hierarchical classification approach presented in this chapter), the remaining 12 columns are di-

vided into four groups of three columns for standard classification, CGUSD, SARSET and hierarchical classification. With respect to standard classification, CGUSD and hierarchical classification their respective columns show: (i) the best algorithm, (ii) the overall accuracy expressed as a percentage (Acc) and (iii) the Area Under the ROC Curve (AUC). With respect to the columns for SARSET the columns show: (i) the part of the data set used when the best results were produced (“F” - First half, “S” - Second half or “E” - Entire data set) indicated by the letter “P”, (ii) the overall accuracy expressed as a percentage (Acc) and (iii) the Area Under the ROC Curve (AUC). Note that for the algorithms that had the best performance with respect to hierarchical classification the hierarchical strategy used is indicated, cascading (“casc”) or not cascading (“-casc”). Where a data set was not used for evaluation purpose with respect to a particular technique this is indicated in the table using a “*” character.

Table 7.8: Overall classification results for the proposed approaches.

Data set	Standard Classification				CGUSD				SARSET				Hierarchical Classification		
	Best Algorithm	Acc (%)	AUC		Best Algorithm	Acc (%)	AUC		P	Acc (%)	AUC		Best Algorithm	Acc (%)	AUC
SAVSNET-840-4-FT	SMO	89.29	0.95		C4.5	55.12	0.75		S	88.86	0.82		*	*	*
SAVSNET-840-4-TD+FT	SMO	90.60	0.96		*	*	*		*	*	*		*	*	*
SAVSNET-971-3-FT	SMO	74.87	0.80		RIPPER	62.62	0.60		F	76.13	0.82		*	*	*
SAVSNET-971-3-TD+FT	SMO	75.30	0.80		*	*	*		*	*	*		*	*	*
SAVSNET-917-1L	SMO	70.34	0.75		RIPPER	61.61	0.61		S	74.36	0.78		SMO (casc/-casc)	70.34	0.75
SAVSNET-917-2L	RIPPER	66.96	0.57		*	*	*		F	75.42	0.59		SMO (-casc)	87.41	0.56
SAVSNET-917-3L	C4.5	57.25	0.59		*	*	*		S	71.66	0.74		SMO (casc)	66.11	0.60
SAVSNET-917-4L	SMO	47.00	0.64		*	*	*		E	52.97	0.07		SMO (-casc)	45.55	0.61
OHSUMED-CA-3187-1L	SMO	94.84	0.93		SMO	89.10	0.90		S	77.21	0.87		SMO (casc/-casc)	94.84	0.93
OHSUMED-CA-2570-2L	SMO	80.82	0.94		RIPPER	73.91	0.90		F	67.78	0.35		SMO (-casc)	86.27	0.93
OHSUMED-CA-834-3L	SMO	78.42	0.90		C4.5	66.88	0.87		S	82.19	0.61		SMO (-casc)	87.69	0.81
OHSUMED-AD-3393-1L	SMO	82.46	0.85		SMO	49.02	0.87		S	84.55	0.65		*	*	*
OHSUMED-AD-569-2L	C4.5	76.74	0.87		SMO	63.33	0.89		F	78.84	0.57		*	*	*
OHSUMED-AD-292-3L	C4.5	89.69	0.91		C4.5	85.57	0.96		S	83.32	0.75		*	*	*
Reuters-21578-LOC-2327-1L	SMO	82.25	0.93		SMO	41.30	0.84		F	79.88	0.59		SMO (casc/-casc)	82.25	0.93
Reuters-21578-LOC-2327-2L	SMO	74.73	0.88		SMO	33.91	0.88		F	77.24	0.52		RIPPER (-casc)	72.62	0.74
Reuters-21578-COM-2327-1L	SMO	95.79	0.98		RIPPER	74.90	0.91		F	91.25	0.88		SMO (casc/-casc)	95.79	0.98
Reuters-21578-COM-2327-2L	SMO	84.40	0.97		*	*	*		F	85.85	0.70		C4.5 (-casc)	81.37	0.87

From Table 7.8 it can be observed that for standard classification, CGUSD and hierarchical classification the algorithm that produced best overall performance was SMO. With respect to the SARSET methodology, for most data sets, the accuracy results obtained were similar with those obtained using SMO for the other approaches. Again it can be seen that the approaches applicable to most of the data sets were standard classification and SARSET.

Note that while almost all the results obtained using standard classification techniques consistently produced good accuracy and AUC results, when CGUSD was applied there were many cases in which low accuracy and high AUC results were obtained for the same data set, which is interesting considering that most of the secondary data sets were balanced (it also demonstrated once again that by using AUC as an evaluation measure it is possible to have more insight into the obtained results than when only accuracy is used). The observations about the results shown in Table 7.8 are considered independently with respect to each of the main data collections because the text contained in each data collection has a different nature. In the following list the performance of the proposed approaches with respect to the different data collections used is presented:

1. **SAVSNET:** Regarding the SAVSNET data sets, while the best accuracy results were obtained using SARSET, the worst accuracy results were obtained using CGUSD. The accuracy results using standard and hierarchical classification were similar with those obtained using SARSET. In terms of AUC the results obtained using standard classification, SARSET and hierarchical classification were very similar and better than those obtained with CGUSD. These considerations are based on the cases where it was possible to make a comparison. The free text from the SAVSNET questionnaires seemed to benefit from straightforward classification techniques and from the provided domain experts knowledge through the use of the semi-automated tool (SARSET). From the classification results shown in Table 7.8 it is concluded that for most of the SAVSNET data sets it was possible to generate informative and relatively good quality summaries using standard classification, SARSET and hierarchical classification.
2. **OHSUMED:** It was possible to use the OHSUMED data sets for evaluating almost all the proposed approaches with the exception of the hierarchical classification approach, for which the OHSUMED-AD-3393-4H hierarchy of classes was not used because of the unbalanced distribution of documents per class and the small number of documents with respect to some classes. Interestingly, the overall results obtained with respect to all the techniques when they were applied to the OHSUMED data sets overcome the results obtained when all the techniques were applied to the SAVSNET data sets, showing that the type of text

used influences the outcome of the approaches used. It can be conjectured that the main reason for obtaining higher results than the SAVSNET data sets is that the free text from the OSHUMED data sets tended to not feature many of the data quality issues present in the SAVSNET data (lack of structure, misspellings, poor grammar, use of abbreviations and acronyms). With respect to the data sets in the OHSUMED-CA-3187-4H hierarchy of classes, the best results in terms of both accuracy and AUC were obtained using standard classification, CGUSD and hierarchical classification. On the other hand the SARSET results were similar in terms of accuracy but not good in terms of AUC. With respect to the data sets in the OHSUMED-AD-3393-4H hierarchy of classes, the best overall results were obtained using the standard classification approach. However, in terms of accuracy, SARSET also obtained good results; and in terms of AUC CGUSD produced the best results.

3. **Reuters-21578:** In the case of the Reuters-21578 data sets it was possible to use almost all of the data sets with respect to the proposed approaches except the Reuters-21578-COM-2327-2L when using CGUSD, because it was difficult to generate secondary data for this data set. Similar to the results obtained using the OHSUMED data sets, the results obtained for the Reuters-21578 data sets were better than those obtained with respect to the SAVSNET data set (for the same conjectured reason). In terms of accuracy the best results were obtained using standard classification, SARSET and hierarchical classification. In terms of AUC the best results were obtained using standard classification, CGUSD and hierarchical classification. Note that the difference found with respect to the accuracy and AUC results obtained when using CGUSD for the OHSUMED data set was more evident for the Reuters-21578 data sets by having in general low accuracy and high AUC results.

As described in detail in Chapter 2, the summaries generated were evaluated in terms of both the intrinsic and extrinsic evaluation measures typically used for text summarisation. Recall that intrinsic evaluation is directed at the analysis and comparison of the generated summary with the original document or with a summary generated by a human. Extrinsic evaluation is directed at determining how useful a summary is with respect to a certain domain. For all the approaches proposed, the generated summaries complied with all the intrinsic evaluation criteria as they were: grammatically correct, non-redundant, complete, accurate, structured and coherent and presented referential clarity. With respect to the extrinsic evaluation measures and considering that the type of text from which it was desired to generate summaries was that from questionnaires, the generated summaries were seen as acceptable taking into account that the results were obtained with text classification, which was the only quantitative evaluation mea-

sure used. In other words, in terms of the extrinsic evaluation measures the quality of the generated summaries depended on how well the classification methods performed.

With respect to the data sets used in this thesis, the best techniques for generating summaries from questionnaires were: (i) standard classification and (ii) SARSET. With respect to medical abstracts, the best techniques for generating summaries were: (i) standard classification and (ii) hierarchical classification. Note that SARSET performed well with respect to AUC. Finally, the best techniques for generating summaries from news items were: (i) standard classification and (ii) hierarchical classification.

With respect to the objectives identified in the introduction to this chapter:

1. *To determine whether the quality of the desired text summarisation classifiers can be improved by benefitting from a hierarchical text classification approach.* From the results obtained it was determined that the quality of the text summarisation classifiers benefit from the hierarchical text classification approach presented in this chapter because it results in a more effective form of classification/summarisation by supporting the incorporation of specialised classifiers at specific nodes in the hierarchy. The use of the class labels for each level of the hierarchy as part of the summaries result in longer and more comprehensive summaries with respect to the other approaches presented in this thesis.
2. *To ascertain the applicability and effectiveness of the approach when applied to free text from different sources than questionnaires.* The applicability and effectiveness of the hierarchical approach presented in this chapter with respect to free text from different sources than questionnaires was demonstrated. It was conjectured that what makes a hierarchy of classes suitable for text summarisation/classification using this approach is not the type of free text, but: (i) the way in which the hierarchy of classes is defined and (ii) the availability of sufficient documents for each level in the hierarchy.
3. *To compare the performance of the summarisation classifiers having hierarchies of classes defined using different criteria (sequential questions, categories of diseases and location and topics of news stories).* As can be seen from Table 7.7, the best results were obtained using SMO, which is a SVM algorithm. It can be conjectured that SMO had the best performance for most of the levels in the hierarchies because its SVM nature allowed it to effectively reduce a multi-class problem into many binary classification problems for each level. SVM are known for obtaining good results with respect to binary classification and have been shown to work well in the context of text mining [112].
4. *To compare the operation of the approach proposed in this chapter with those of the previous chapters and to report any improvements with respect to the classification*

results obtained in Chapters 4, 5 and 6. In order to compare the operation of the hierarchical approach proposed in this chapter with those of the previous chapters it was decided to use averaged values for accuracy and AUC as the best way to make comparisons, as it was shown in Table 7.8. In general classification results were better than those obtained with CGUSD and they were similar with those obtained using standard classification techniques and with SARSET. An overall comparison of the proposed approaches was presented.

7.6 Summary

This chapter has presented a hierarchical classification approach aimed at the generation of text summarisation classifiers. Two hierarchical classification strategies were devised in terms of the scope and dependency between the levels in the hierarchies used: (i) cascading and (ii) non-cascading. While the former took into account the output of the classifier at the parent nodes with respect to the child nodes at the next level in the hierarchy, the latter generated classifiers independently from the output of parent nodes. As in previous chapters, the hierarchical approach presented in this chapter was evaluated with different types of free text: (i) free text from questionnaires, (ii) free text from medical abstracts and (iii) free text from news stories from a news agency. In order to conduct a comparison of the performance of the hierarchical approaches with respect to the approaches presented in Chapters 4, 5 and 6, the same standard flat classification evaluation measures were used, namely: (i) overall accuracy expressed as a percentage, (ii) Area Under the ROC Curve (AUC), (iii) precision, (iv) sensitivity/recall and (v) specificity. The comparison of the performance of all the approaches presented in this thesis was presented.

Overall good results were obtained considering: (i) the unbalanced nature of the data sets used and (ii) that in some cases it was not possible to associate sufficient numbers of records with particular classes (nodes) in the hierarchies. In terms of the hierarchical strategies carried out, and not taking into account the root nodes of the hierarchies (which of course produced the same classification results regardless of which strategy was used), the results were similar for both the cascading and the non-cascading strategies. Classification results were better than those obtained with CGUSD and they were similar with those obtained using standard classification and SARSET. It was not possible to compare the operation of the hierarchical approach when applied to the OHSUMED-AD-3393-3H hierarchy of classes. However, in general, it is argued that more comprehensive summaries were produced with respect to the CGUSD and SARSET approaches due to the inclusion of a greater number of class labels.

Chapter 8

Conclusions and Future Work

This chapter provides a summary of the proposed text classification methods used for text summarisation that were presented in this thesis, the main findings and contributions, and some possible future directions. Section 8.1 presents the summary of the proposed approaches in terms of their objectives and operation. The main findings and contributions of the research presented in this thesis is presented in Section 8.2. Finally the potential directions for possible future research are presented in Section 8.3.

8.1 Summary

The aim of the research presented in this thesis, as stated in Chapter 1, was to answer the following research question: *“Is it possible to generate summaries describing the free text element often found in questionnaires using text classification techniques; while at the same time taking into account that such text is usually sparse, unstructured and contains misspelled words, poor grammar, and abbreviations and acronyms related to a specific domain?”*. Five research issues associated with this research question were identified:

1. The inherent characteristics of the free text part of questionnaires.
2. The need for robust classification techniques whose results have an effect on the quality of the generated summaries.
3. The mechanism for generating the desired summaries from class labels.
4. The typical case of having unbalanced data sets.
5. The requirement for sufficient training data so that effective questionnaire classification summarisation techniques can be applied.

Recall that this thesis has presented a study on the use of text classification methods for text summarisation with respect to the unstructured free text part of questionnaire data. At the same time the proposed techniques have also been applied and evaluated

to other forms of free text. The fundamental idea presented in this thesis is that text summarisation can be conceived of as a form of text classification in that the classes assigned to text documents may be viewed as indicators (summarisers) of the main concepts contained in the original free text, but in a coherent and reduced form. Coherent because the class names that are typically used to label text documents tend to represent a synthesis of the topic with which the document is concerned. Reduced because the context has been minimised to a set of labels. It was acknowledged that a summary of this form was not as complete or as extensive as what some practitioners might consider to be a summary; however by assigning multiple class labels to each document (questionnaire free text) the generated summaries, it is argued, come close to what might be traditionally viewed as a summary.

Chapter 2 presented a review of the relevant background knowledge with respect to the research addressed in this thesis. Three main areas were covered: (i) Questionnaire Data Mining, (ii) Text Classification and (iii) Text Summarisation. Chapter 3 presented a description of the nature of the data sets used to evaluate the proposed approaches as well as how they were preprocessed. Three data collections containing different types of text were considered: (i) SAVSNET (questionnaire free text), (ii) OHSUMED (text from medical abstracts) and (iii) Reuters-21578 (text from news stories). Overall 18 subsets of these data collections were used for evaluation purposes. The number of data sets actually used for evaluation purposes with respect to the proposed techniques in each case varied depending on the nature of the technique. The motivation for using different types of text, other than free text from questionnaires, was so as to demonstrate the wide applicability and effectiveness of the presented approaches although they were initially intended for questionnaire data.

Four approaches were presented to address the aforementioned research question:

1. **Using Standard Classification Techniques for Text Summarisation.** Chapter 4 considered the use of standard classification techniques for text summarisation. The motivation was to establish a benchmark with which the more specialised summarisation classification techniques presented later in the thesis could be later compared. A number of different classifier generators were used. Since the questionnaire data also contained a tabular data part, this was also considered in order to determine whether the inclusion of tabular data in the classification process might improve the effectiveness of the classification results and consequently the quality of the summaries. Two feature selection techniques were also considered: (i) Term Frequency-Inverse Document Frequency (TF-IDF) plus Chi-squared and (ii) TF-IDF plus Correlation-based Feature Selection (CFS).
2. **Classifier Generation Using Secondary Data (CGUSD) for Text Summarisation.** Chapter 5 presented an approach that considered the case when

the available data was not considered sufficient for training purposes (or possibly because no data was available at all). The idea was to consider building the desired text summarisation classifiers using some appropriate secondary data set and then applying the resulting classifier, for the purposes of text summarisation, to the primary data (free text that is to be summarised). A challenge with respect to the proposed CGUSD technique was the necessity to first obtain appropriate secondary data that featured the same topics (class labels) as the primary data. The main objective was to determine whether text summarisation classifiers could be generated using secondary data which could then be effectively applied to primary data to produce good quality summaries. The issue of having unbalanced data sets for generating a classifier was partially addressed by CGUSD because it provided for the possibility of generating balanced secondary data sets (regardless of whether the primary data was balanced or not). The operation of the classifiers that had the best performance in Chapter 4, using standard classification techniques, was compared with respect to the CGUSD approach. In terms of accuracy, for all the data sets in which both approaches could be applied, CGUSD had the worst performance. In terms of AUC almost all the compared results showed that standard classification performed better than the CGUSD approach.

3. **Using a Semi-Automated Rule Summarisation Extraction Tool (SARSET) for Text Summarisation.** Chapter 6 presented a semi-automated classification technique called SARSET (Semi-Automated Rule Summarisation Extraction Tool) to support document summarisation classification. The motivation for SARSET was the conjecture that the results obtained with respect to the approaches presented in Chapters 4 and 5 could be improved upon by allowing a domain expert to take part in the process for generating summarisation classifiers. SARSET allowed domain experts to select phrases from questionnaire returns (or other types of text) in a training set that could be included in the antecedents of classification rules. For each phrase selected by the domain expert SARSET automatically generated variations of the suggested phrases, using a synonym database and “wild card” characters. The domain expert then identified the most relevant phrase variations within the context of the document set and a set of classification rules was produced where the antecedents comprised the selected phrase variations and the consequent was the name of the class (*phrase variation* \Rightarrow *name of class*). Exceptions, in this context, were specific phrases that could be covered by a rule antecedent, but which should not be used for classification purposes. As with the previous approaches SARSET was evaluated using different types of text in order to ascertain its applicability and effectiveness. Overall in terms of accuracy it performed similarly to the standard classification approach and better than CGUSD. In terms of AUC SARSET

performed worst than the standard classification approach and the CGUSD approach.

4. **Text Summarisation Using Hierarchical Text Classification.** Chapter 7 presented a hierarchical summarisation classification approach. This approach assumed that text summarisation could be achieved using a classification approach whereby several class labels could be associated with documents which then constituted the summarisation. This hierarchical summarisation classification approach offered the advantage that different levels of classification could be used and the summarisation could be customised according to which branch of the tree the current document was located. Hierarchical classification involves the use of class labels arranged in a tree structure, therefore hierarchical text classification is a form of multi-label classification. The advantages offered by the proposed hierarchical summarisation classification approach were: (i) people are familiar with the concept of defining things in a hierarchical manner, (ii) hierarchies are a good way of encapsulating knowledge, (iii) not all the class labels have to be taken into account for summarising unseen records and (iv) hierarchical classification summarisation is more effective than summarisation based on flat classification approaches given that specialised classifiers are used in each node of the hierarchy of classes. Unlike the approaches presented in Chapters 4, 5 and 6, the use of hierarchical classification allowed the use of more than one class label to be used to generate summaries, thus the summaries produced were nearer to what might be traditionally viewed as a summary. The main objective was to determine if the quality of the desired text summarisation classifiers could be improved by using the hierarchical text classification approach. Overall, in the cases where it was possible to make a comparison, in terms of accuracy the hierarchical text classification approach produced a better performance than the standard classification, CGUSD and SARSET approaches; in terms of AUC the hierarchical classification approach performed similarly to the standard classification approach and better than the CGUSD and SARSET approaches.

A detailed description of how the research issues were addressed is presented in the next section.

8.2 Main Findings and Contributions

This thesis focused on the summarisation of free text found in questionnaires using text classification methods. Four summarisation approaches, founded on the concept of text classification, were proposed: (i) using standard classification techniques, (ii) using secondary data to generate classifiers to be applied to primary data, (iii) using a semi-automated rule summarisation extraction tool that requires user interaction, and (iv)

using a hierarchical text classification approach to generate the desired summaries. The design and operation of these approaches addressed the research question introduced in Chapter 1 and a number of associated research issues. In this section the research question and associated research issues, presented in Chapter 1 and repeated in the introduction to this chapter, are returned to.

Recall that the research question was:

“can relevant information be extracted in the form of a summary from the free text element often found in questionnaires using text classification techniques; while at the same time taking into account that such text is usually sparse, unstructured and contains misspelled words, poor grammar, and abbreviations and acronyms related to a specific domain?”.

The work described in this thesis indicates that the answer to this question is that summaries can be extracted from the free text element often found in questionnaires using appropriately defined text classification techniques. Furthermore, it is argued, the applicability and effectiveness of the proposed approaches have been demonstrated.

With respect to the five research issues associated with the provision to the answer of the research question the following main findings were arrived at:

1. **Inherent characteristics of the free text part of questionnaires.** Recall that these inherent characteristics were: (i) a lack of structure, (ii) misspelled words, (iii) poor grammar and (iv) the use of abbreviations and acronyms. With respect to the lack of structure it was considered early on in this research to use a VSM representation, more specifically the bag-of-words representation whereby free text is “tokenised” using white space characters and delimiter symbols in order to isolate words and store them in a vector (one vector per document). In the bag-of-words representation the relationship between words is lost, but this is not considered to be a disadvantage with respect to questionnaire free text which tends to have very little structure anyway. On the contrary it allows for faster computational processing. Stemming was used to overcome the presence of misspelled words and poor grammar in the free text. Stemming also served to reduce the size of the feature space. Misspellings were no longer an issue if they occurred in the part of the word that derived from the stems. The use of abbreviations and acronyms was addressed by the standard classification approach, the CGUSD approach and the hierarchical classification approach by including those abbreviations and acronyms that appeared more frequently in the stop words list; with respect to SARSET, the knowledge of domain experts helped to identify relevant abbreviations and acronyms from the free text.
2. **Robust classification techniques.** The robustness of the proposed approaches was demonstrated because they were satisfactorily applied to a range of data sets

used for evaluation purposes in this thesis, which featured between 2 and 92 classes. The largest data sets considered were: (i) OHSUMED-AD-3393-1L which featured 3,393 documents and was used to evaluate the standard classification and the SARSET techniques, (ii) the secondary data generated with respect to the OHSUMED-CA-2570-2L which featured 5,535 documents and was used to evaluate the CGUSD technique and (iii) OHSUMED-CA-3187-1L which featured 3,187 documents and was used to evaluate the hierarchical classification technique.

3. **The mechanism for generating the desired summaries from class labels.**

As detailed in Chapter 4, the mechanism for generating the desired summaries relies on: (i) the resulting class labels assigned to unseen documents as a result of applying a selected classifier and (ii) simple rules established by domain experts to decide how to prepend or append domain-specific text to the generated class labels. The text to be prepended or appended is based entirely on the domain expert's criteria and in the context of the application domain of the free text. Regarding the summary generation of the approaches presented in Chapters 4, 5 and 6, only one label per document was considered; which, although this resulted in very concise summaries, were short compared to what many practitioners consider as a summary. The hierarchical nature of the approach presented in Chapter 7 allowed the use of more class labels. Generating summaries from free text in this manner was deemed to be effective in the particular case where text is sparse, unstructured, contains misspelled words, poor grammar, and abbreviations related to a specific domain, such as the free text part of questionnaires.

4. **Unbalanced data.** Recall that unbalanced data refers either to the case where a significant number of documents per class is not available or to the case where there is a large number of documents with respect to some classes but not others. This typically occurs in real-world data sets. Almost all of the data sets used in this thesis to evaluate the performance of the proposed approaches were unbalanced. It is argued that the presence of unbalanced data was the most significant issue effecting performance with respect to the techniques proposed in this thesis. In Chapter 5 this issue was partially addressed using the CGUSD approach by generating balanced secondary data when there was not sufficient primary data. Interestingly, while balanced secondary data was used for generating classifiers in Chapter 5 using the CGUSD approach, the AUC results obtained were better than the ones obtained using both SARSET and hierarchical classification, but worse than those obtained when standard classification techniques were applied directly to unbalanced data sets. In terms of accuracy, CGUSD produced the worst performance. Note that AUC gives a better insight into performance than accuracy in the presence of unbalanced data. In general the obtained results for

all the approaches are considered as satisfactory considering the unbalanced nature of the data sets, demonstrating the applicability of the proposed approaches to unbalanced data sets.

5. **Sufficient training data.** Having sufficient training data was another issue identified early on in this research, especially with respect to the questionnaire data sets. To address this issue the CGUSD approach was proposed where appropriate secondary data was extracted from source data closely related to the primary data in order to build the desired text summarisation classifiers. With respect to the data sets used for evaluation purposes, in most cases, it was possible to generate sufficient secondary data. The reasons why it was not possible in all cases was that the classes in the candidate secondary data sets were not always found to be entirely compatible with the classes defined in the primary data sets.

The main contributions of the work described in this thesis are as follows:

1. An evaluation of the use of the concept of text classification in the context of questionnaire free text summarisation.
2. A demonstration of the benefits of the usage of classification for summarisation that addresses the issues associated with the nature of the free text element of questionnaire data, which tends to be unstructured and include misspellings, abbreviations and domain specific terminology.
3. An investigation into the use of secondary data to support free text classification and specifically questionnaire free text summarisation.
4. A hierarchical classification mechanism to provide more sophisticated text summarisation, with respect to questionnaire data, than that provided using single label classifiers.
5. A number of approaches to generating summaries from the free text element of questionnaire data, namely: (i) standard classification techniques, (ii) use of secondary data, (iii) semi-automated summary generation (requires end user involvement) and (iv) hierarchical classification for text summarisation.
6. An investigation into the SAVSNET [83] questionnaire data collections.

8.3 Future Directions

The research described in this thesis has indicated a number of promising directions for future research. These research directions are briefly outlined below.

1. **Alternative integration of tabular data and free text from questionnaires.** In the approach that used standard classification techniques presented in Chapter 4 relevant features from the tabular data part of questionnaires were added to the free text part as keywords in an attempt to add more relevant information to the free text and therefore to improve the classification results. However, as the experiments carried out using the SAVSNET-840-4-TD+FT and the SAVSNET-971-3-TD+FT data sets demonstrated, the inclusion of tabular data in the form of keywords did not improve the quality of the summarisation classification. One alternative way to integrate tabular data with free text data so as to improve the summarisation classification of free text would be by using the most relevant tabular attributes directly with respect to the generated summaries. Another way would be to use the spatial and temporal tabular attributes along with the other most relevant tabular attributes in order to provide more context in relation to the generated summaries.
2. **Experiments with alternative data sets.** The approaches presented in this thesis to investigate the use of classification methods for text summarisation were focused mainly on the free text part of questionnaires. However two other types of free text were used to evaluate the proposed approaches besides questionnaire text, namely: (i) text from medical abstracts and (ii) text from news stories. The motivation for using different types of free text was to demonstrate the applicability and effectiveness of the approaches when applied to free text from different sources than questionnaires. In order to widen the scope of the approaches presented in this thesis more extensive experiments need to be carried out using: (i) different types of data from that used in the experiments presented in this thesis such as “tweets” (text messages of up to 140 characters which people mainly use to share thoughts or to give opinions about topics) from the Twitter social network, opinions of people regards videos on YouTube or comments of people about news in on-line editions of newspapers, and (ii) different forms of questionnaire free text data such as that frequently included in the “end of module” feedback reports that are typically completed by university students or product review questionnaires.
3. **Including multi-label classification.** With the exception of the hierarchical classification approach presented in Chapter 7, all the approaches used single-label classification techniques to generate the summarisation classifiers. While it was found that using a single label produced concise, coherent and informative summaries from documents, it is expected that using more class labels (as demonstrated using the proposed hierarchical classification technique) will enrich the generated summaries. However, finding data sets where more labels are included

and then creating the mechanisms for multi-label classification with respect to text summarisation is not a straightforward task. Therefore it is suggested that future work is required so as to comprehensively address this issue. With respect to the hierarchical classification approach, it is suggested that experiments using larger hierarchies (more than four levels) should also be carried out.

4. **Finding more efficient mechanisms to address unbalanced data.** Handling unbalanced data sets was partially addressed in Chapter 5 using the CGUSD approach and, while the obtained AUC results were better than those obtained using SARSET and the hierarchical classification approaches, they were worse than those produced using standard classification. In terms of accuracy CGUSD produced the worst performance. The results presented in this thesis demonstrated that: (i) the presence of unbalanced data is almost inevitable in the context of real-world data and (ii) generating classifier from balanced data sets (as has been suggested by some practitioners) does not necessarily produce satisfactory results. Therefore finding more efficient mechanisms to address the unbalanced data issue is considered to be an important item for future research.
5. **SARSET improvements.** The SARSET technique allowed for the use of domain experts knowledge to improve the classification based summarisation of free text from questionnaires. It was demonstrated that SARSET could be effectively applied to free text from different sources than questionnaire data and that the results obtained were similar to those obtained with respect to the other approaches presented in this thesis. However, in the context of the operation of SARSET there are many things that can be improved such as: (i) the way in which the rules are generated and “fired” during the classification process, (ii) the GUI interface and (iii) the performance optimisation of the proposed techniques. It is therefore suggested that this will provide another fertile direction for future work.
6. **Use of Data Stream Mining.** The techniques presented in this thesis could include or address Data Stream Mining in the context of Questionnaire Data Mining because, as it was indicated in the introductory chapter of this thesis, two of the motivations for this work are the volume of data to be analysed and the increase in the speed of how the data is analysed. In the case of the standard classification technique for example, Figure 4.1 would have a loop to represent the continuous flow of incoming data.

Bibliography

- [1] A. Abd-Elrahman, M. Andreu, and T. Abbott. Using text data mining techniques for understanding free-style question answers in course evaluation forms. *Research in Higher Education Journal*, 9:12–23, 2010.
- [2] S. Afantenos, V. Karkaletsis, and P. Stamatopoulos. Summarization from medical documents: a survey. *Artificial Intelligence in Medicine*, 33(2):157–177, 2005.
- [3] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, volume 1215, pages 487–499. Citeseer, 1994.
- [4] L. Alonso, I. Castellón, S. Climent, M. Fuentes, L. Padró, and H. Rodríguez. Approaches to text summarization: Questions and answers. *Inteligencia Artificial*, 8:22, 2004.
- [5] Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas. Using coreference chains for text summarization. In *CorefApp '99: Proceedings of the Workshop on Coreference and its Applications*, pages 77–84, Morristown, NJ, USA, 1999. Association for Computational Linguistics.
- [6] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York., 1999.
- [7] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, volume 17. Madrid, Spain, 1997.
- [8] M. Bramer. *Principles of Data Mining*. Springer London Ltd, Published, 2007.
- [9] A. Celikyilmaz and D. Hakkani-Tür. Concept-based classification for multi-document summarization. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5540–5543. IEEE, 2011.
- [10] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni. Hierarchical classification: combining bayes with svm. In *Proceedings of the 23rd international conference on Machine learning*, pages 177–184. ACM, 2006.

- [11] C.C. Chang and C.J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.
- [12] N.V. Chawla. Data mining for imbalanced datasets: An overview. *Data Mining and Knowledge Discovery Handbook*, pages 875–886, 2010.
- [13] Hao Chen and Tin K Ho. Evaluation of decision forests on text categorization. In *Electronic Imaging*, pages 191–199. International Society for Optics and Photonics, 1999.
- [14] Y.L. Chen and C.H. Weng. Mining fuzzy association rules from questionnaire data. *Knowledge-Based Systems*, 22(1):46–56, 2009.
- [15] Wesley T. Chuang and Jihoon Yang. Extracting sentence segments for text summarization: a machine learning approach. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 152–159, New York, NY, USA, 2000. ACM.
- [16] F. Coenen. Lucs-kdd dn software (version 2), 2003.
- [17] Frans Coenen. The LUCS-KDD TFP association rule mining algorithm. http://www.csc.liv.ac.uk/frans/KDD/Software/Apriori_TFP/aprioriTFP.html, 2004.
- [18] Frans Coenen. The LUCS-KDD TFPC classification association rule mining algorithm. http://www.csc.liv.ac.uk/frans/KDD/Software/Apriori_TFPC/aprioriTFPC.html, 2004.
- [19] William W. Cohen. Fast effective rule induction. In Armand Prieditis and Stuart Russell, editors, *Proc. of the 12th International Conference on Machine Learning*, pages 115–123, Tahoe City, CA, July 9–12, 1995. Morgan Kaufmann.
- [20] Eduardo P Costa, Ana C Lorena, André CPLF Carvalho, Alex A Freitas, and Nicholas Holden. Comparing several approaches for hierarchical classification of proteins with decision trees. In *Advances in Bioinformatics and Computational Biology*, pages 126–137. Springer, 2007.
- [21] Tomaz Curk, Janez Demsar, Qikai Xu, Gregor Leban, Uros Petrovic, Ivan Bratko, Gad Shaulsky, and Blaz Zupan. Microarray data mining with visual programming. *Bioinformatics*, 21(3):396–398, 2005.
- [22] D. Das and A.F.T. Martins. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II Course at CMU*, 2007.
- [23] S. Deerwester. Improving information retrieval with latent semantic indexing. 1988.

- [24] S. DAlessio, M. Murray, R. Schiaffino, and A. Kershenbaum. Category levels in hierarchical text categorization. In *Proceedings of EMNLP-3, 3rd conference on empirical methods in natural language processing*. sn, 1998.
- [25] A. Edmunds and A. Morris. The problem of information overload in business organisations: a review of the literature. *International journal of information management*, 20(1):17–28, 2000.
- [26] Gonenc Ercan and Ilyas Cicekli. Using lexical chains for keyword extraction. *Information Processing & Management*, 43(6):1705 – 1714, 2007. Text Summarization.
- [27] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, et al. Knowledge discovery and data mining: Towards a unifying framework. In *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR*, pages 82–88, 1996.
- [28] D. Fragoudis, D. Meretakos, and S. Likothanassis. Best terms: an efficient feature-selection algorithm for text categorization. *Knowledge and Information Systems*, 8(1):16–33, 2005.
- [29] M. Fuentes and H. Rodríguez. Using cohesive properties of text for automatic summarization. *JOTRI02*, 2002.
- [30] M. F. Garcia-Constantino, F. Coenen, P. Noble, and A. Radford. Questionnaire free text summarisation using hierarchical classification. In *AI-2012: Research and Development in Intelligent Systems XXIX. Incorporating Applications and Innovations in Intelligent Systems XX.*, pages 35–48. Springer, 2012.
- [31] M. F. Garcia-Constantino, F. Coenen, P. Noble, and A. Radford. Free text summarisation of structured and unstructured free text using hierarchical classification. 2013.
- [32] M. F. Garcia-Constantino, F. Coenen, P. Noble, A. Radford, and C. Setzkorn. A semi-automated approach to building text summarisation classifiers. In Petra Perner, editor, *Eight International Conference on Machine Learning and Data Mining*, pages 495–509. Springer, 2012.
- [33] M. F. Garcia-Constantino, F. Coenen, P. Noble, A. Radford, and C. Setzkorn. A semi-automated approach to building text summarisation classifiers. *Journal of Theoretical and Applied Computer Science*, 6(4):7–23, 2012.
- [34] M. F. Garcia-Constantino, F. Coenen, P. Noble, A. Radford, C. Setzkorn, and A. Tierney. An investigation concerning the generation of text summarisation classifiers using secondary data. In Petra Perner, editor, *Seventh International*

- Conference on Machine Learning and Data Mining*, pages 387–398. Springer, 2011.
- [35] B. Gillham. *Developing a questionnaire*. Continuum Intl Pub Group, 2000.
 - [36] Yihong Gong and Xin Liu. Generic text summarization using relevance measure and latent semantic analysis. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25, New York, NY, USA, 2001. ACM.
 - [37] M. Gotoh, T. Sakai, J. Itoh, T. Ishida, and S. Hirasawa. Knowledge discovery from questionnaires with items and texts (in japanese). *Proceedings of 2003 PC Conference, Kagoshima, Japan*, August 2003.
 - [38] Michael Granitzer. Hierarchical text classification using methods from machine learning. Master’s thesis, 2003.
 - [39] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.
 - [40] M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.
 - [41] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, second edition edition, 2005.
 - [42] D.J. Hand and R.J. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001.
 - [43] H. Hardy, N. Shimizu, T. Strzalkowski, L. Ting, X. Zhang, and G.B. Wise. Cross-document summarization by concept classification. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 121–128. ACM, 2002.
 - [44] W. Hersh, C. Buckley, TJ Leone, and D. Hickam. Ohsumed: an interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 192–201. Springer-Verlag New York, Inc., 1994.
 - [45] A. Hiramatsu, H. Oiso, S. Tamura, and N. Komoda. Support system for analyzing open-ended questionnaires data by culling typical opinions. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 2, pages 1377–1382. IEEE, 2004.

- [46] S. Hirasawa. Analyses of student questionnaires for faculty developments. 2006.
- [47] S. Hirasawa and W.W. Chu. Knowledge acquisition from documents with both fixed and free formats. In *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, volume 5, pages 4694–4699. IEEE, 2003.
- [48] S. Hirasawa, T. Ishida, H. Adachi, M. Gotoh, and T. Sakai. A document classification and its application to questionnaire analyses (in japanese). *Proceedings of 2005 Spring Conference in Information Management, JASMIN, Tokyo.*, pages 54–57, June 2005.
- [49] S. Hirasawa, T. Ishida, J. Itoh, M. Gotoh, and T. Sakai. Analyses on student questionnaires with fixed and free formats (in japanese). *Proceedings of Computer Education JUCE*, pages 144–145, September 2003.
- [50] I. Hiroko, U. Masao, and I. Hitoshi. Criterion for judging request intention in response texts of open-ended questionnaires. In *Proceedings of the second international workshop on Paraphrasing-Volume 16*, pages 49–56. Association for Computational Linguistics, 2003.
- [51] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.
- [52] A. Hotho, A. Nürnberger, and G. Paaß. A brief survey of text mining. In *LDV Forum-GLDV Journal for Computational Linguistics and Language Technology*, volume 20, pages 19–62. Citeseer, 2005.
- [53] Phoenix X Huang, Bastiaan J Boom, and Robert B Fisher. Hierarchical classification for live fish recognition.
- [54] H. Inui and H. Isahara. Proposition for "extended modality" - extraction of intention in open-ended response texts. *Technical Report of EICE, NLC2002*, 102(414):31–36, 2002.
- [55] H. Inui, M. Murata, K. Uchimoto, and H. Isahara. Classification of open-ended questionnaires based on surface information in sentence structure. In *Proceedings of the 6th NLPRS2001*, pages 315–322, 2001.
- [56] H. Inui, K. Uchimoto, M. Murata, and H. Isahara. Classification of open-ended questionnaires based on predicative (in japanese). *NLP*, 99:181–188, 1998.
- [57] T. Ishida, M. Gotoh, and S. Hirasawa. Analysis of student questionnaire in the lecture of computer science (in japanese). *Computer Education, CIEC*, 18:152–159, July 2005.

- [58] M. Jaoua and A. Hamadou. Automatic text summarization of scientific articles based on classification of extracts population. *Computational Linguistics and Intelligent Text Processing*, pages 363–377, 2003.
- [59] H. Jing. Sentence reduction for automatic text summarization. In *Proceedings of the 6th Applied Natural Language Processing Conference*, pages 310–315, 2000.
- [60] H. Jing and K.R. McKeown. Cut and paste based text summarization. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, pages 178–185. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2000.
- [61] T. Joachims. *Learning to classify text using support vector machines: Methods, theory and algorithms*, volume 186. Kluwer Academic Publishers Norwell, MA, USA:, 2002.
- [62] Thorsten Joachims. Making large scale svm learning practical. 1999.
- [63] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.
- [64] K.S. Jones et al. Automatic summarizing: factors and directions. *Advances in automatic text summarization*, pages 1–12, 1999.
- [65] A.K. Joshi. Natural language processing. *Science*, 253(5025):1242, 1991.
- [66] Byung Soo Kim, Jae Young Park, Anush Mohan, Anna Gilbert, and Silvio Savarese. Hierarchical classification of images by sparse approximation. In *British machine vision conference*. Citeseer, 2011.
- [67] K. Knight and D. Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107, 2002.
- [68] David D Lewis, Robert E Schapire, James P Callan, and Ron Papka. Training algorithms for linear text classifiers. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 298–306. ACM, 1996.
- [69] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.
- [70] Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958.

- [71] Inderjeet Mani. Recent developments in text summarization. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 529–531, New York, NY, USA, 2001. ACM.
- [72] G. Marshall. The purpose, design and administration of a questionnaire for data collection. *Radiography*, 11(2):131–136, 2005.
- [73] R. Mihalcea. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)(companion volume)*, 2004.
- [74] G. Miner, J. Elder IV, T. Hill, R. Nisbet, and D. Delen. *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press, 2012.
- [75] Tom Mitchell, Terry Russell, Peter Broomhead, and Nicola Aldridge. Towards robust computerised marking of free-text responses. 2002.
- [76] J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1):21–48, 1991.
- [77] Y. Murakami, Y. Tanizawa, D. Han, and M. Harada. Automatic classification of open-ended questionnaires based on semantic analysis (in japanese). *The 66th National Convention of IPSJ*, pages 171–172, 2005.
- [78] M. Nagamachi. Kansei engineering: a new ergonomic consumer-oriented technology for product development. *International Journal of industrial ergonomics*, 15(1):3–11, 1995.
- [79] T. Nasukawa. Text mining application for call centers. *Journal of the Japanese Society for Artificial Intelligence*, 16(2):219–225, 2001.
- [80] J.C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.
- [81] J.R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.
- [82] J.R. Quinlan. *C4.5: Programs for Machine Learning*, volume 1. Morgan Kaufmann, 1993.
- [83] A. Radford, Á. Tierney, KP Coyne, RM Gaskell, PJ Noble, S. Dawson, C. Setzkorn, PH Jones, IE Buchan, JR Newton, et al. Developing a network for small animal disease surveillance. *Veterinary Record*, 167(13):472–474, 2010.
- [84] M. Rogati and Y. Yang. High-performing feature selection for text classification. In *Proceedings of the eleventh international conference on Information and knowledge management*, page 661. ACM, 2002.

- [85] R.J. Roiger and M.W. Geatz. *Data Mining: A tutorial-based primer*. Addison Wesley Boston, 2003.
- [86] M. Rosell and S. Velupillai. Revealing relations between open and closed answers in questionnaires through text clustering evaluation. In *Proc. of the Sixth Int. Conf. on Language Resources and Evaluation (LREC08)*, 2008.
- [87] T. Sakai, T. Ishida, M. Gotoh, and S. Hirasawa. A student questionnaires analysis system based on natural language expressions (in japanese). *IEICE*, 2004.
- [88] T. Sakai, J. Itoh, M. Gotoh, T. Ishida, and S. Hirasawa. Efficient analysis of student questionnaires using information retrieval techniques (in japanese). *Proceedings of 2003 Spring Conference on Information Management, JASMIN*, pages 182–185, June 2003.
- [89] M. Saravanan, P.C.R. Raj, and S. Raman. Summarization and categorization of text data in high-level data cleaning for information retrieval. *Applied Artificial Intelligence*, 17(5-6):461–474, 2003.
- [90] F. Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.
- [91] C.E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.
- [92] H. Gregory Silber and Kathleen F. McCoy. Efficient text summarization using lexical chains. In *IUI '00: Proceedings of the 5th international conference on Intelligent user interfaces*, pages 252–255, New York, NY, USA, 2000. ACM.
- [93] H.G. Silber and K.F. McCoy. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics*, 28(4):487–496, 2002.
- [94] Abraham Silberschatz and Alexander Tuzhilin. On subjective measures of interestingness in knowledge discovery. In *KDD*, volume 95, pages 275–281, 1995.
- [95] Carlos N Silla Jr and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2):31–72, 2011.
- [96] Stephen Soderland. Learning to extract text-based information from the world wide web. In *KDD*, volume 97, pages 251–254, 1997.
- [97] J. Steinberger and K. Ježek. Text summarization and singular value decomposition. *Lecture Notes in Computer Science*, pages 245–254, 2004.

- [98] J. Steinberger and K. Ježek. Evaluation measures for text summarization. *Computing and Informatics*, 28(2):251–275, 2012.
- [99] A. Sun and E.P. Lim. Hierarchical text classification and evaluation. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 521–528. IEEE, 2001.
- [100] V. Svátek. Ontologies, questionnaires and (mining) tabular data. In *the 3rd European Semantic Web Conference (ESWC 2006)*, 2006.
- [101] K. Takahashi. A supporting system for cording of the answers from open-ended question. *Sociological Theory and Methods*, 15(1):149–164, 2000.
- [102] M. Tateno. The method to extract textual ”kansei” expression in the customer’s voice. *IPSJ SIG Notes, NL*, 153(14):105–112, 2003.
- [103] K. Toutanova, F. Chen, K. Popat, and T. Hofmann. Text classification in a hierarchical mixture model for small training sets. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 105–113. ACM, 2001.
- [104] Y. Uchida, T. Yoshikawa, T. Furuhashi, E. Hirao, and H. Iguchi. Extraction of important keywords in free text of questionnaire data and visualization of relationship among sentences. In *Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference on*, pages 1604–1608. IEEE, 2009.
- [105] Y.J. Wang. *Language-independent pre-processing of large documentbases for text classification*. PhD thesis, The University of Liverpool, 2007.
- [106] P. Willett. The porter stemming algorithm: then and now. *Program: electronic library and information systems*, 40(3):219–223, 2006.
- [107] I.H. Witten, E. Frank, and M.A. Hall. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.
- [108] K. Yamanishi and H. Li. Mining open answers in questionnaire data. *IEEE Intelligent Systems*, pages 58–63, 2002.
- [109] T. Yanase, S. Marumoto, I. Nanba, and R. Ochitani. Parsing question texts using the predicate expressions of the sentence end. *Proceedings of the 8th Annual Meeting of the Association for NLP*, pages 647–650, 2002.
- [110] Y. Yang. An evaluation of statistical approaches to medline indexing. In *Proceedings of the AMIA Annual Fall Symposium*, page 358. American Medical Informatics Association, 1996.

- [111] Y. Yang and J.O. Pedersen. A comparative study on feature selection in text categorization. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 412–420. Citeseer, 1997.
- [112] Waleed Zaghloul, Sang M Lee, and Silvana Trimi. Text classification: neural networks vs support vector machines. *Industrial Management & Data Systems*, 109(5):708–717, 2009.
- [113] Z. Zheng, X. Wu, and R. Srihari. Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explorations Newsletter*, 6(1):80–89, 2004.

Appendix A

Distribution of records per class with respect to the evaluation data sets used

In this appendix the distribution of records per class with respect to the evaluation data sets is presented (Tables A.1 to A.14). For the tables related to the SAVSNET (Tables A.1 to A.6) and Reuters-21578 (Tables A.13 to A.16) data sets, the columns are as follows: (i) the first column shows the name of the class, (ii) the second column shows the number of records per class and (iii) the third column shows the percentage of records per class. For the tables related to the OHSUMED (Tables A.7 to A.12) data sets, the columns are explained as follows: (i) the first column shows the name of the class, (ii) the second column shows the MeSH Tree Code related to the class, (iii) the third column shows the number of records per class and (iv) the fourth column shows the percentage of records per class. Note that the last row of all the tables gives the total number and percentage of records in the data set.

Table A.1: Number of records per class in SAVSNET-840-4 and in SAVSNET-840-4-TD+FT ($k = 352$).

Class	Num.	%
<i>Aggression</i>	34	4.05
<i>Diarrhoea</i>	315	37.50
<i>Pruritus</i>	352	41.90
<i>Vomiting</i>	139	16.55
Total	840	100.00

Table A.2: Number of records per class in SAVSNET-971-3 and in SAVSNET-971-3-TD+FT ($k = 586$).

Class	Num.	%
<i>Diarrhoea</i>	586	60.35
<i>Vomiting</i>	268	27.60
<i>Vom & Dia</i>	117	12.05
Total	971	100.00

Table A.3: Number of records per class in SAVSNET-917-1L ($k = 536$).

Class	Num.	%
<i>Diarrhoea</i>	536	58.45
<i>Vomiting</i>	248	27.05
<i>Vom & Dia</i>	133	14.50
Total	917	100.00

Table A.4: Number of records per class in SAVSNET-917-2L ($k = 604$).

Class	Num.	%
<i>Haemorrhagic</i>	177	19.30
<i>Not Haemorrhagic</i>	604	65.87
<i>Unknown Severity</i>	136	14.83
Total	917	100.00

Table A.5: Number of records per class in SAVSNET-917-3L ($k = 573$).

Class	Num.	%
<i>First Time</i>	573	62.49
<i>Nth Time</i>	290	31.62
<i>Unknown Occurrence</i>	54	5.89
Total	917	100.00

Table A.6: Number of records per class in SAVSNET-917-4L ($k = 411$).

Class	Num.	%
<i>Less Than One Day</i>	273	29.77
<i>Between Two And Four Days</i>	411	44.82
<i>Between Five And Seven Days</i>	82	8.94
<i>More Than Eight Days</i>	139	15.16
<i>Unknown Duration</i>	12	1.31
Total	917	100.00

Table A.7: Number of records per class in OHSUMED-CA-3187-1L ($k = 2,339$).

Class	MeSH Tree Code	Num.	%
<i>Congenital Heart Defects</i>	C14.240.400	2,339	73.39
<i>Vascular Malformations</i>	C14.240.850	848	26.71
Total		3,187	100.00

Table A.8: Number of records per class in OHSUMED-CA-2570-2L ($k = 525$).

Class	MeSH Tree Code	Num.	%
<i>Cor Triatriatum</i>	C14.240.400.200	21	0.82
<i>Coronary Vessel Anomalies</i>	C14.240.400.210	218	8.48
<i>Crisscross Heart</i>	C14.240.400.220	6	0.23
<i>Dextrocardia</i>	C14.240.400.280	11	0.43
<i>Patent Ductus Arteriosus</i>	C14.240.400.340	160	6.23
<i>Eisenmenger Complex</i>	C14.240.400.450	22	0.86
<i>Heart Septal Defects</i>	C14.240.400.560	524	20.39
<i>Levocardia</i>	C14.240.400.701	2	0.08
<i>Marfan Syndrome</i>	C14.240.400.725	106	4.12
<i>Noonan Syndrome</i>	C14.240.400.787	16	0.62
<i>Tetralogy of Fallot</i>	C14.240.400.849	153	5.95
<i>Aortic Coarctation</i>	C14.240.400.90	237	9.22
<i>Transposition of Great Vessels</i>	C14.240.400.915	227	8.83
<i>Arteriovenous Malformations</i>	C14.240.850.750	525	20.43
<i>Scimitar Syndrome</i>	C14.240.850.968	13	0.51
<i>Vascular Fistula</i>	C14.240.850.984	329	12.80
Total		2,570	100.00

Table A.9: Number of records per class in OHSUMED-CA-834-3L ($k = 299$).

Class	MeSH Tree Code	Num.	%
<i>Endocardial Cushion Defects</i>	C14.240.400.560.350	21	2.52
<i>Atrial Heart Septal Defects</i>	C14.240.400.560.375	191	22.90
<i>Ventricular Heart Septal Defects</i>	C14.240.400.560.540	228	27.34
<i>Aortopulmonary Septal Defect</i>	C14.240.400.560.98	34	4.08
<i>Double Outlet Right Ventricle</i>	C14.240.400.915.300	31	3.72
<i>Arterio – Arterial Fistula</i>	C14.240.850.984.500	30	3.60
<i>Arteriovenous Fistula</i>	C14.240.850.984.750	299	35.85
Total		834	100.00

Table A.10: Number of records per class in OHSUMED-AD-3393-1L ($k = 2,112$).

Class	MeSH Tree Code	Num.	%
<i>Bird Diseases</i>	C22.131	65	1.92
<i>Borna Disease</i>	C22.152	6	0.18
<i>Cat Diseases</i>	C22.180	52	1.53
<i>Cattle Diseases</i>	C22.196	87	2.56
<i>Veterinary Abortion</i>	C22.21	7	0.21
<i>Animal Disease Models</i>	C22.232	2,112	62.25
<i>Dog Diseases</i>	C22.268	118	3.48
<i>Erysipelothrix Infections</i>	C22.331	5	0.15
<i>Fish Diseases</i>	C22.362	15	0.44
<i>Foot – and – Mouth Disease</i>	C22.380	8	0.24
<i>Actinobacillosis</i>	C22.39	1	0.03
<i>Foot Rot</i>	C22.394	2	0.06
<i>Goat Diseases</i>	C22.405	5	0.15
<i>Heartwater Disease</i>	C22.434	1	0.03
<i>Animal Hepatitis</i>	C22.467	82	2.42
<i>Horse Diseases</i>	C22.488	34	1.00
<i>Infectious Keratoconjunctivitis</i>	C22.500	1	0.03
<i>Animal Lameness</i>	C22.510	1	0.03
<i>Animal Muscular Dystrophy</i>	C22.595	76	2.24
<i>Aleutian Mink Disease</i>	C22.62	1	0.03
<i>Animal Parasitic Diseases</i>	C22.674	236	6.96
<i>Paratuberculosis</i>	C22.688	9	0.27
<i>Parturient Paresis</i>	C22.695	1	0.03
<i>Contagious Pleuropneumonia</i>	C22.717	1	0.03
<i>Primate Diseases</i>	C22.735	89	2.62
<i>Pseudorabies</i>	C22.742	5	0.15
<i>Rinderpest</i>	C22.780	5	0.15
<i>Rodent Diseases</i>	C22.795	39	1.15
<i>Animal Salmonella Infections</i>	C22.812	29	0.85
<i>Sheep Diseases</i>	C22.836	110	3.24
<i>Swine Diseases</i>	C22.905	23	0.68
<i>Veterinary Venereal Tumors</i>	C22.950	1	0.03
<i>Vesicular Stomatitis</i>	C22.952	68	2.00
<i>Zoonoses</i>	C22.969	98	2.89
Total		3,393	100.00

Table A.11: Number of records per class in OHSUMED-AD-569-2L ($k = 194$).

Class	MeSH Tree Code	Num.	%
<i>Newcastle Disease</i>	C22.131.630	13	2.28
<i>Poultry Diseases</i>	C22.131.728	21	3.69
<i>Avian Tuberculosis</i>	C22.131.921	2	0.35
<i>Avian Leukosis</i>	C22.131.94	20	3.51
<i>Feline Acquired Immunodeficiency Syndrome</i>	C22.180.350	6	1.05
<i>Bovine Virus Diarrhea – Mucosal Disease</i>	C22.196.106	2	0.35
<i>Freemartinism</i>	C22.196.339	1	0.18
<i>Infectious Bovine Rhinotracheitis</i>	C22.196.429	5	0.88
<i>Bovine Tuberculosis</i>	C22.196.927	3	0.53
<i>Distemper</i>	C22.268.265	13	2.28
<i>Canine Hip Dysplasia</i>	C22.268.485	1	0.18
<i>Furunculosis</i>	C22.362.224	5	0.88
<i>Animal Viral Hepatitis</i>	C22.467.435	63	11.07
<i>Equine Infectious Anemia</i>	C22.488.304	3	0.53
<i>Animal Helminthiasis</i>	C22.674.377	42	7.38
<i>Animal Protozoan Infections</i>	C22.674.710	194	34.09
<i>Monkey Diseases</i>	C22.735.500	89	15.64
<i>Infectious Ectromelia</i>	C22.795.239	1	0.18
<i>Murine Acquired Immunodeficiency Syndrome</i>	C22.795.650	5	0.88
<i>Border Disease</i>	C22.836.160	1	0.18
<i>Contagious Ecthyma</i>	C22.836.259	8	1.41
<i>Progressive Interstitial Pneumonia of Sheep</i>	C22.836.660	6	1.05
<i>Scrapie</i>	C22.836.799	49	8.61
<i>Visna</i>	C22.846.900	14	2.46
<i>Gastroenteritis Transmissible of Swine</i>	C22.905.469	1	0.18
<i>African Swine Fever</i>	C22.905.72	1	0.18
Total		569	100.00

Table A.12: Number of records per class in OHSUMED-AD-292-3L ($k = 133$).

Class	MeSH Tree Code	Num.	%
<i>Rift Valley Fever</i>	C22.467.435.812	23	7.88
<i>Dirofilariasis</i>	C22.674.377.320	19	6.51
<i>Toxocariasis</i>	C22.674.377.868	23	7.88
<i>Babesiosis</i>	C22.674.710.122	19	6.51
<i>Cryptosporidiosis</i>	C22.674.710.235	133	45.55
<i>Theileriasis</i>	C22.674.710.735	6	2.05
<i>Animal Toxoplasmosis</i>	C22.674.710.817	36	12.33
<i>Marburg Virus Disease</i>	C22.735.500.500	3	1.03
<i>Simian Acquired Immunodeficiency Syndrome</i>	C22.735.500.850	30	10.27
Total		292	100.00

Table A.13: Number of records per class in Reuters-21578-LOC-2327-1L ($k = 1,021$).

Class	Num.	%
<i>AfricaAmerica</i>	51	2.19
<i>AfricaAsia</i>	7	0.30
<i>Africa</i>	70	3.01
<i>AfricaEurope</i>	19	0.82
<i>AmericaAsia</i>	243	10.44
<i>AmericaAustralia</i>	6	0.26
<i>America</i>	1,021	43.88
<i>AmericaEurope</i>	90	3.87
<i>AsiaAustralia</i>	7	0.30
<i>Asia</i>	314	13.49
<i>AsiaEurope</i>	100	4.30
<i>Australia</i>	40	1.72
<i>AustraliaEurope</i>	2	0.09
<i>Europe</i>	357	15.34
Total	2,327	100.00

Table A.14: Number of records per class in Reuters-21578-COM-2327-1L ($k = 966$).

Class	Num.	%
<i>Energy</i>	633	27.20
<i>Grains</i>	743	31.93
<i>Livestock</i>	105	4.51
<i>Metal</i>	347	14.91
<i>Soft</i>	499	21.44
Total	2,327	100.00

Table A.15: Number of records per class in Reuters-21578-LOC-2327-2L ($k = 743$).

Class	Num.	%
<i>Algeria</i>	4	0.17
<i>Argentina</i>	24	1.03
<i>Australia</i>	38	1.63
<i>Austria</i>	3	0.13
<i>Bahrain</i>	4	0.17
<i>Bangladesh</i>	7	0.30
<i>Belgium</i>	60	2.58
<i>Bolivia</i>	6	0.26
<i>Botswana</i>	1	0.04
<i>Brazil</i>	54	2.32
<i>Bulgaria</i>	1	0.04
<i>Burma</i>	1	0.04
<i>Cameroon</i>	2	0.09
<i>Canada</i>	133	5.72
<i>Chile</i>	2	0.09
<i>China</i>	40	1.72
<i>Colombia</i>	22	0.95
<i>Costa Rica</i>	3	0.13
<i>Cuba</i>	5	0.21
<i>Cyprus</i>	6	0.26
<i>Denmark</i>	2	0.09
<i>Dominican Republic</i>	1	0.04
<i>Ecuador</i>	21	0.90
<i>Egypt</i>	5	0.21
<i>El Salvador</i>	1	0.04
<i>Ethiopia</i>	1	0.04
<i>Fiji</i>	2	0.09
<i>Finland</i>	5	0.21
<i>France</i>	54	2.32
<i>Ghana</i>	5	0.21
<i>Greece</i>	6	0.26
<i>Haiti</i>	4	0.17
<i>Hong Kong</i>	3	0.13
<i>Hungary</i>	1	0.04
<i>India</i>	17	0.73
<i>Indonesia</i>	45	1.93
<i>Iran</i>	10	0.43
<i>Iraq</i>	7	0.30
<i>Ireland</i>	1	0.04
<i>Israel</i>	1	0.04
<i>Italy</i>	8	0.34
<i>Ivory Coast</i>	6	0.26
<i>Japan</i>	79	3.39
<i>Jordan</i>	1	0.04

Table continues in the next page.

Table A.15 continued.

Class	Num.	%
<i>Kenya</i>	5	0.21
<i>Kuwait</i>	9	0.39
<i>Luxembourg</i>	5	0.21
<i>Madagascar</i>	4	0.17
<i>Malaysia</i>	24	1.03
<i>Mauritius</i>	2	0.09
<i>Mexico</i>	9	0.39
<i>Netherlands</i>	31	1.33
<i>New Zealand</i>	2	0.09
<i>Nicaragua</i>	3	0.13
<i>Nigeria</i>	3	0.13
<i>Norway</i>	8	0.34
<i>Pakistan</i>	11	0.47
<i>Papua New Guinea</i>	1	0.04
<i>Peru</i>	9	0.39
<i>Philippines</i>	12	0.52
<i>Poland</i>	2	0.09
<i>Portugal</i>	2	0.09
<i>Qatar</i>	2	0.09
<i>Saudi Arabia</i>	15	0.64
<i>Singapore</i>	11	0.47
<i>South Africa</i>	17	0.73
<i>South Korea</i>	5	0.21
<i>Spain</i>	10	0.43
<i>Sri Lanka</i>	10	0.43
<i>Suriname</i>	1	0.04
<i>Sweden</i>	4	0.17
<i>Switzerland</i>	23	0.99
<i>Syria</i>	1	0.04
<i>Taiwan</i>	19	0.82
<i>Tanzania</i>	5	0.21
<i>Thailand</i>	25	1.07
<i>Togo</i>	1	0.04
<i>Trinidad Tobago</i>	1	0.04
<i>Turkey</i>	3	0.13
<i>UAE</i>	8	0.34
<i>Uganda</i>	5	0.21
<i>UK</i>	247	10.61
<i>USA</i>	966	41.51
<i>USSR</i>	21	0.90
<i>Venezuela</i>	21	0.90
<i>Vietnam</i>	1	0.04
<i>West Germany</i>	42	1.80
<i>Yemen Arab Republic</i>	3	0.13
<i>Yemen Demo Republic</i>	1	0.04
<i>Yugoslavia</i>	2	0.09

Table continues in the next page.

Table A.15 continued.

Class	Num.	%
<i>Zambia</i>	7	0.30
<i>Zimbabwe</i>	6	0.26
Total	2,327	100.00

Table A.16: Number of records per class in Reuters-21578-COM-2327-2L ($k = 508$).

Class	Num.	%
<i>Alum</i>	48	2.06
<i>Barley</i>	1	0.04
<i>Carcass</i>	29	1.25
<i>Cocoa</i>	60	2.58
<i>Coconut</i>	2	0.09
<i>Coffee</i>	124	5.33
<i>Copper</i>	61	2.62
<i>Corn</i>	8	0.34
<i>Cotton</i>	27	1.16
<i>Crude</i>	486	20.89
<i>F – cattle</i>	2	0.09
<i>Fishmeal</i>	1	0.04
<i>Fuel</i>	13	0.56
<i>Gas</i>	33	1.42
<i>Gold</i>	120	5.16
<i>Grain</i>	508	21.83
<i>Groundnut</i>	3	0.13
<i>Heat</i>	16	0.69
<i>Hog</i>	16	0.69
<i>Iron – steel</i>	51	2.19
<i>Jet</i>	4	0.17
<i>L – cattle</i>	2	0.09
<i>Lead</i>	19	0.82
<i>Livestock</i>	56	2.41
<i>Lumber</i>	13	0.56
<i>Meal – feed</i>	22	0.95
<i>Naphtha</i>	1	0.04
<i>Nat – gas</i>	48	2.06
<i>Nickel</i>	5	0.21
<i>Oilseed</i>	77	3.31
<i>Orange</i>	21	0.90
<i>Palm – oil</i>	3	0.13
<i>Pet – chem</i>	29	1.25
<i>Platinum</i>	4	0.17
<i>Plywood</i>	2	0.09
<i>Potato</i>	5	0.21
<i>Propane</i>	3	0.13
<i>Rapeseed</i>	2	0.09

Table continues in the next page.

Table A.16 continued.

Class	Num.	%
<i>Rice</i>	3	0.13
<i>Rubber</i>	41	1.76
<i>Silver</i>	16	0.69
<i>Soy – meal</i>	1	0.04
<i>Soybean</i>	4	0.17
<i>Strategic – meal</i>	19	0.82
<i>Sugar</i>	146	6.27
<i>Tapioca</i>	1	0.04
<i>Tea</i>	9	0.39
<i>Tin</i>	32	1.38
<i>Veg – oil</i>	88	3.78
<i>Wheat</i>	21	0.90
<i>Wool</i>	1	0.04
<i>Zinc</i>	20	0.86
Total	2,327	100.00

Appendix B

Additional Experimental Results

This appendix presents experimental results used for the evaluation of the hierarchical classification approach presented in Chapter 7 which were omitted from the main body of this thesis because of space restrictions. The additional experimental results are shown from Tables B.1 to B.15. For all the tables the first row indicates which classification algorithm was used (SMO, C4.5 or RIPPER), the second row indicates the hierarchical classification approach that was used (cascading or non-cascading) and the third row is the header of the results table. With respect to the latter row, the first column presents the level of the hierarchy and the second column presents the nodes from which the results were obtained. The remaining ten columns are divided in two groups of five columns according to the hierarchical classification strategy applied (cascading and non-cascading respectively); for each group the results obtained are presented in terms of the evaluation measures used: (i) overall accuracy expressed as a percentage (Acc), (ii) Area Under the ROC Curve (AUC), (iii) precision (Pr), (iv) sensitivity/recall (Sn/Re) and (v) specificity (Sp).

Table B.1: Classification results for SAVSNET-917-4H using SMO.

Algorithm Approach	Node	SMO									
		casc					\neg casc				
		Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
First	GI Diseases	70.34	0.75	0.68	0.70	0.75	70.34	0.75	0.68	0.70	0.75
Second	Diarrhoea	64.01	0.62	0.61	0.64	0.59	74.63	0.68	0.74	0.75	0.61
	Vomiting Vomiting & Diarrhoea	75.43 55.21	0.52 0.60	0.68 0.56	0.75 0.55	0.29 0.63	89.11 98.50	0.49 0.50	0.82 0.97	0.89 0.99	0.10 0.02
Third	Haemorrhagic Not Haemorrhagic Unknown Severity	69.17 61.10 68.06	0.64 0.60 0.58	0.63 0.59 0.64	0.69 0.61 0.68	0.58 0.57 0.47	61.02 63.08 58.82	0.59 0.59 0.61	0.57 0.59 0.57	0.61 0.63 0.59	0.56 0.53 0.62
Fourth	First Time	51.18	0.62	0.49	0.51	0.66	47.99	0.59	0.44	0.48	0.66
	Nth Time Unknown Occurrence	37.34 *	0.60 *	0.36 *	0.37 *	0.75 *	43.10 51.85	0.64 0.46	0.38 0.39	0.43 0.52	0.75 0.40

Table B.2: Classification results for SAVSNET-917-4H using C4.5.

Algorithm Approach	Node	C4.5									
		casc					\neg casc				
		Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
First	GI Diseases	63.14	0.69	0.61	0.63	0.71	63.14	0.69	0.61	0.63	0.71
Second	Diarrhoea	61.20	0.58	0.58	0.61	0.54	70.90	0.65	0.70	0.71	0.55
	Vomiting Vomiting & Diarrhoea	61.83 46.43	0.60 0.49	0.58 0.45	0.62 0.46	0.53 0.57	87.10 98.50	0.47 0.10	0.81 0.97	0.87 0.99	0.09 0.02
Third	Haemorrhagic Not Haemorrhagic Unknown Severity	60.76 54.87 54.35	0.57 0.54 0.55	0.60 0.52 0.55	0.61 0.55 0.54	0.57 0.53 0.54	50.85 58.61 41.91	0.52 0.55 0.47	0.48 0.57 0.44	0.51 0.59 0.42	0.52 0.52 0.60
Fourth	First Time	42.68	0.57	0.41	0.43	0.68	39.27	0.53	0.39	0.39	0.67
	Nth Time Unknown Occurrence	30.77 *	0.52 *	0.31 *	0.31 *	0.73 *	30.34 50.00	0.54 0.52	0.31 0.51	0.30 0.50	0.73 0.59

Table B.3: Classification results for SAVSNET-917-4H using RIPPER.

		RIPPER											
Algorithm		casc						¬casc					
Approach													
Level	Node	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp		
First	GI Diseases	67.18	0.69	0.63	0.67	0.67	67.18	0.69	0.63	0.67	0.67		
Second	Diarrhoea	70.23	0.63	0.63	0.70	0.50	76.87	0.65	0.76	0.77	0.51		
	Vomiting	63.43	0.55	0.52	0.63	0.43	89.92	0.47	0.82	0.90	0.10		
Third	Vomiting & Diarrhoea	27.78	0.42	0.28	0.28	0.53	98.50	0.10	0.97	0.99	0.02		
	Haemorrhagic	59.38	0.42	0.41	0.59	0.36	55.37	0.49	0.44	0.55	0.43		
	Not Haemorrhagic	60.33	0.51	0.52	0.60	0.42	64.40	0.53	0.57	0.64	0.40		
Fourth	Unknown Severity	62.50	0.59	0.42	0.63	0.47	58.09	0.58	0.53	0.58	0.54		
	First Time	46.59	0.57	0.44	0.47	0.63	46.42	0.53	0.40	0.46	0.57		
	Nth Time	36.67	0.44	0.19	0.37	0.56	40.34	0.56	0.35	0.40	0.69		
	Unknown Occurrence	*	*	*	*	*	61.11	0.58	0.58	0.61	0.56		

Table B.4: Classification results for OHSUMED-CA-3187-3H using SMO.

Algorithm Approach	Level	Node	SMO									
			casc					¬casc				
			Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
First Second	Types of CA Congenital Heart Defects Vascular Malformations		94.84	0.93	0.95	0.95	0.91	94.84	0.93	0.95	0.95	0.91
			81.36	0.93	0.83	0.81	0.94	79.95	0.94	0.81	0.80	0.96
Third	Heart Septal Defects Vascular Fistula		89.01	0.88	0.89	0.89	0.87	92.58	0.92	0.92	0.93	0.91
			77.81	0.83	0.78	0.78	0.83	80.98	0.88	0.75	0.81	0.92
			96.36	0.56	0.97	0.96	0.15	94.40	0.73	0.94	0.94	0.52

Table B.5: Classification results for OHSUMED-CA-3187-3H using C4.5.

Algorithm Approach	Level	Node	C4.5									
			casc					¬casc				
			Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
First	Types of CA	Congenital Heart Defects	93.16	0.91	0.93	0.93	0.86	93.16	0.91	0.93	0.93	0.86
			79.27	0.88	0.80	0.79	0.94	75.01	0.88	0.75	0.75	0.96
Second		Vascular Malformations	90.52	0.90	0.91	0.91	0.89	93.89	0.93	0.94	0.94	0.93
Third	Heart Septal Defects	Vascular Fistula	78.21	0.83	0.78	0.78	0.86	71.17	0.82	0.72	0.71	0.91
			93.69	0.63	0.94	0.94	0.45	94.00	0.76	0.94	0.94	0.61

Table B.6: Classification results for OHSUMED-CA-3187-3H using RIPPER.

Algorithm Approach	Level	Node	RIPPER									
			casc					¬casc				
			Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
First	Types of CA	Congenital Heart Defects	93.38	0.90	0.93	0.93	0.86	93.38	0.90	0.93	0.93	0.86
			83.25	0.91	0.84	0.83	0.95	76.51	0.90	0.79	0.77	0.95
Second		Vascular Malformations	91.80	0.91	0.92	0.92	0.91	91.27	0.89	0.90	0.91	0.90
Third	Heart Septal Defects	Vascular Fistula	82.17	0.86	0.84	0.82	0.89	74.23	0.82	0.71	0.74	0.88
			95.98	0.49	0.95	0.96	0.28	94.00	0.75	0.94	0.94	0.61

Table B.7: Classification results for OHSUMED-AD-3393-3H using SMO.

Algorithm Approach	Level	Node	SMO									
			casc					\neg casc				
			Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
Second	Types of AD	Bird Diseases	82.46	0.85	0.81	0.83	0.77	82.46	0.85	0.81	0.83	0.77
		Cattle Diseases	*	0.86	0.91	0.90	0.82	*	*	*	*	*
		Dog Diseases	*	*	*	*	*	14.29	0.40	0.18	0.14	0.79
		Animal Parasitic Diseases	*	*	*	*	*	92.86	0.50	0.86	0.93	0.07
		Rodent Diseases	*	*	*	*	*	95.87	0.89	0.96	0.96	0.82
		Sheep Diseases	*	*	*	*	*	*	*	*	*	*
		Swine Diseases	*	*	*	*	*	84.85	0.80	0.83	0.85	0.78
Third	Types of AD	Animal Helminthiasis	*	*	*	*	*	53.85	0.60	0.49	0.54	0.70
		Animal Protozoan Infections	*	*	*	*	*	92.61	0.89	0.93	0.93	0.85

Table B.8: Classification results for OHSUMED-AD-3393-3H using C4.5.

Algorithm Approach	Level	Node	C4.5									
			casc					\neg casc				
			Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
Second	Types of AD	Bird Diseases	79.89	0.81	0.77	0.80	0.79	79.89	0.81	0.77	0.80	0.79
		Cattle Diseases	70.83	0.78	0.68	0.71	0.74	*	*	*	*	*
		Dog Diseases	*	*	*	*	*	21.43	0.40	0.13	0.21	0.70
		Animal Parasitic Diseases	*	*	*	*	*	92.86	0.04	0.86	0.93	0.07
		Rodent Diseases	*	*	*	*	*	92.20	0.80	0.93	0.92	0.66
		Sheep Diseases	*	*	*	*	*	*	*	*	*	*
		Swine Diseases	*	*	*	*	*	77.27	0.76	0.77	0.77	0.83
Third	Types of AD	Animal Helminthiasis	*	*	*	*	*	57.69	0.68	0.50	0.58	0.78
		Animal Protozoan Infections	*	*	*	*	*	89.77	0.88	0.91	0.90	0.87

Table B.9: Classification results for OHSUMED-AD-3393-3H using RIPPER.

Algorithm Approach	Node	RIPPER									
		casc					\neg casc				
		Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
First	Types of AD	80.71	0.80	0.79	0.81	0.78	80.71	0.80	0.79	0.81	0.78
Second	Bird Diseases	81.48	0.88	0.80	0.82	0.87	*	*	*	*	*
	Cattle Diseases	*	*	*	*	*	35.71	0.25	0.13	0.36	0.64
	Dog Diseases	*	*	*	*	*	92.86	0.04	0.86	0.93	0.07
	Animal Parasitic Diseases	*	*	*	*	*	96.33	0.87	0.97	0.96	0.84
	Rodent Diseases	*	*	*	*	*	*	*	*	*	*
	Sheep Diseases	*	*	*	*	*	75.76	0.79	0.72	0.76	0.80
	Swine Diseases	*	*	*	*	*	*	*	*	*	*
Third	Animal Helminthiasis	*	*	*	*	*	53.85	0.62	0.45	0.54	0.73
	Animal Protozoan Infections	*	*	*	*	*	90.34	0.87	0.91	0.90	0.85

Table B.10: Classification results for Reuters-21578-LOC-2327-2H using SMO.

		SMO											
Algorithm													
Approach													
Level		casc						¬casc					
	Node	Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp		
First	Regions												
Second	AfricaAmerica	82.25	0.93	0.81	0.82	0.94	82.25	0.93	0.81	0.82	0.94		
	AfricaAsia	*	*	*	*	*	31.37	0.30	*	*	*		
	Africa	73.08	0.84	0.71	0.73	0.95	69.57	0.85	0.70	0.70	0.93		
	AfricaEurope	*	*	*	*	*	68.42	0.59	0.50	0.68	0.77		
	AmericaAsia	68.98	0.70	0.58	0.69	0.64	73.53	0.71	0.64	0.74	0.63		
	AmericaAustralia	*	*	*	*	*	*	*	*	*	*		
	America	87.41	0.84	0.81	0.87	0.79	92.92	0.92	0.92	0.93	0.87		
	AmericaEurope	58.49	0.60	0.45	0.59	0.80	58.89	0.72	0.56	0.59	0.80		
	AsiaAustralia	*	*	*	*	*	*	*	*	*	*		
	Asia	71.73	0.87	0.70	0.72	0.96	76.64	0.92	0.78	0.77	0.96		
	AsiaEurope	71.01	0.54	0.59	0.71	0.51	65.98	0.60	0.55	0.66	0.65		
	Australia	87.88	0.45	0.77	0.88	0.12	87.50	0.47	0.77	0.88	0.13		
	AustraliaEurope	*	*	*	*	*	*	*	*	*	*		
	Europe	76.70	0.85	0.70	0.77	0.89	77.97	0.87	0.75	0.78	0.88		

Table B.11: Classification results for Reuters-21578-LOC-2327-2H using C4.5.

Algorithm	Approach	Node	C4.5									
			casc					\neg casc				
			Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
First		Regions	73.79	0.83	0.72	0.74	0.92	73.79	0.83	0.72	0.74	0.92
Second		AfricaAmerica	77.97	0.73	0.70	0.78	0.70	31.37	0.56	0.27	0.31	0.80
		AfricaAsia	27.27	0.62	0.27	0.27	0.79	*	*	*	*	*
		Africa	27.66	0.64	0.32	0.28	0.93	81.16	0.90	0.83	0.81	0.98
		AfricaEurope	33.33	0.52	0.32	0.33	0.74	42.11	0.60	0.49	0.42	0.79
		AmericaAsia	57.61	0.70	0.54	0.58	0.82	69.33	0.70	0.66	0.69	0.70
		AmericaAustralia	*	*	*	*	*	*	*	*	*	*
		America	83.04	0.84	0.79	0.83	0.85	90.33	0.90	0.89	0.90	0.88
		AmericaEurope	40.00	0.63	0.35	0.40	0.86	54.44	0.76	0.53	0.54	0.91
		AsiaAustralia	*	*	*	*	*	*	*	*	*	*
		Asia	73.20	0.86	0.72	0.73	0.98	85.53	0.93	0.86	0.86	0.99
		AsiaEurope	64.86	0.75	0.56	0.65	0.85	60.82	0.77	0.65	0.61	0.93
		Australia	72.73	0.59	0.66	0.73	0.18	87.50	0.70	0.79	0.88	0.30
		AustraliaEurope	*	*	*	*	*	*	*	*	*	*
		Europe	71.51	0.84	0.70	0.72	0.95	75.94	0.86	0.76	0.76	0.96

Table B.12: Classification results for Reuters-21578-LOC-2327-2H using RIPPER.

Algorithm	Approach	Node	RIPPER									
			casc					\neg casc				
			Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
First		Regions	71.85	0.84	0.70	0.72	0.88	71.85	0.84	0.70	0.72	0.88
Second		AfricaAmerica	88.37	0.61	0.78	0.88	0.62	43.14	0.54	0.34	0.43	0.77
		AfricaAsia	*	*	*	*	*	*	*	*	*	*
		Africa	56.25	0.69	0.49	0.56	0.85	71.01	0.85	0.71	0.71	0.93
		AfricaEurope	*	*	*	*	*	57.89	0.46	0.47	0.58	0.69
		AmericaAsia	72.73	0.74	0.66	0.73	0.75	74.37	0.71	0.68	0.74	0.67
		AmericaAustralia	*	*	*	*	*	*	*	*	*	*
		America	77.79	0.80	0.73	0.78	0.79	88.83	0.87	0.88	0.89	0.85
		AmericaEurope	42.31	0.43	0.20	0.42	0.65	61.11	0.70	0.54	0.61	0.86
		AsiaAustralia	*	*	*	*	*	*	*	*	*	*
		Asia	64.89	0.82	0.63	0.65	0.96	81.25	0.91	0.81	0.81	0.98
		AsiaEurope	67.69	0.64	0.53	0.68	0.69	70.10	0.70	0.54	0.70	0.71
		Australia	80.56	0.31	0.72	0.81	0.31	87.50	0.55	0.77	0.88	0.13
		AustraliaEurope	*	*	*	*	*	*	*	*	*	*
		Europe	71.02	0.79	0.64	0.71	0.86	76.23	0.84	0.73	0.76	0.89

Table B.13: Classification results for Reuters-21578-COM-2327-2H using SMO.

Algorithm Approach	Node	SMO									
		casc					\neg casc				
		Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
First	Types of commodities	95.79	0.98	0.96	0.96	0.99	95.79	0.98	0.96	0.96	0.99
Second	Energy	89.21	0.82	0.87	0.89	0.72	72.20	0.91	0.72	0.72	0.93
	Grains	79.52	0.78	0.72	0.80	0.75	82.73	0.82	0.78	0.83	0.78
	Livestock	70.41	0.72	0.62	0.70	0.78	79.17	0.79	0.80	0.79	0.78
	Metal	79.60	0.90	0.75	0.80	0.94	77.42	0.90	0.75	0.77	0.92
	Soft	89.47	0.94	0.85	0.90	0.97	88.30	0.95	0.89	0.88	0.96

Table B.14: Classification results for Reuters-21578-COM-2327-2H using C4.5.

Algorithm Approach	Node	C4.5									
		casc					\neg casc				
		Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
First	Types of commodities	88.27	0.94	0.88	0.88	0.96	88.27	0.94	0.88	0.88	0.96
Second	Energy	75.82	0.71	0.73	0.76	0.64	70.46	0.85	0.70	0.71	0.97
	Grains	65.81	0.74	0.63	0.66	0.81	78.18	0.79	0.75	0.78	0.80
	Livestock	58.02	0.71	0.52	0.58	0.82	77.78	0.81	0.77	0.78	0.84
	Metal	80.81	0.90	0.78	0.81	0.97	86.29	0.93	0.86	0.86	0.97
	Soft	86.08	0.93	0.85	0.86	0.99	94.15	0.98	0.94	0.94	0.99

Table B.15: Classification results for Reuters-21578-COM-2327-2H using RIPPER.

Algorithm Approach	Node	RIPPER									
		casc					\neg casc				
		Acc (%)	AUC	Pr	Sn/Re	Sp	Acc (%)	AUC	Pr	Sn/Re	Sp
First	Types of commodities	91.41	0.96	0.92	0.91	0.97	91.41	0.96	0.92	0.91	0.97
Second	Energy	83.72	0.77	0.81	0.84	0.71	72.04	0.87	0.69	0.72	0.95
	Grains	68.89	0.72	0.59	0.69	0.73	78.73	0.80	0.74	0.79	0.78
	Livestock	54.39	0.66	0.50	0.54	0.75	70.83	0.69	0.70	0.71	0.73
	Metal	77.81	0.88	0.76	0.78	0.94	77.82	0.90	0.79	0.78	0.93
	Soft	95.26	0.97	0.95	0.95	0.99	94.71	0.97	0.94	0.95	0.98

Appendix C

Published Work

In this final appendix a list of publications to date, including papers submitted for refereeing, from the work described in this thesis is presented:

Refereed Conferences:

1. M. F. Garcia-Constantino, F. Coenen, P. Noble, A. Radford, C. Setzkorn, and A. Tierney. An Investigation Concerning the Generation of Text Summarisation Classifiers using Secondary Data. Proceedings. Machine Learning and Data Mining in Pattern Recognition, Lecture Notes in Computer Science 6871 Springer 2011, pp387-398.
2. S. Chua, F. Coenen, G. Malcolm, and M. F. Garcia-Constantino. Using Negation and Phrases in Inducing Rules for Text Classification. Proceedings. AI-2011: Research and Development in Intelligent Systems XXVIII (Incorporating Applications and Innovations in Intelligent Systems XIX), Springer 2011, pp153-166.
3. M. F. Garcia-Constantino, F. Coenen, P. Noble, A. Radford, and C. Setzkorn. A Semi-Automated Approach to Building Text Summarisation Classifiers. Proceedings. Machine Learning and Data Mining in Pattern Recognition. Lecture Notes in Computer Science. Springer 2012.
4. M. F. Garcia-Constantino, F. Coenen, P. Noble and A. Radford. Questionnaire Free Text Summarisation Using Hierarchical Text Classification. Proceedings. AI-2012: Research and Development in Intelligent Systems XXIX (Incorporating Applications and Innovations in Intelligent Systems XX), Springer 2012, pp35-48.

Journals:

1. M. F. Garcia-Constantino, F. Coenen, P. Noble, A. Radford, and C. Setzkorn. A Semi-Automated Approach to Building Text Summarisation Classifiers. Journal of Theoretical and Applied Computer Science. Vol. 6, No. 4, pp. 7-23, 2012.

Technical Reports:

1. M. F. Garcia-Constantino. Technical report of the text summarisation of the free text Section of questionnaires related to Veterinary practice. 2010.
2. M. F. Garcia-Constantino and F. Coenen. A Survey on Questionnaire Data Mining. 2011.
3. M. F. Garcia-Constantino, F. Coenen, P. Noble and A. Radford. Free Text Summarisation of Structured and Unstructured Free Text Using Hierarchical Classification. 2012.