

AN INTELLIGENT MODEL FOR ONLINE RECRUITMENT FRAUDS DETECTION
BY:

BANDAR MOHAMMED AL - GHAMDI

A THESIS

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF:

MASTER OF SCIENCE IN INFORMATION SECURITY

TO:

INFORMATION SECURITY DEPARTMENT
THE COLLEGE OF COMPUTER AND INFORMATION SECURITY
NAIF ARAB UNIVERSITY FOR SECURITY SCIENCES

MARCH 2018

SUPERVISED BY:

DR. FAHAD AL- HARBY
NAIF ARAB UNIVERSITY FOR SECURITY SCIENCES

Table of Contents

AN INTELLIGENT MODEL FOR ONLINE RECRUITMENT FRAUDS DETECTION	I
Abstract	II
المستخلص	III
Dedication	IV
Acknowledgment	V
Table of Contents	VI
List of Figures	VIII
List of Acronyms	X
Chapter 1	1
Introduction.....	1
1.1 Background	1
1.2 Problem Statement.....	2
1.3 Objectives	3
1.4 Motivation	4
1.5 Contributions	5
1.6 Scope and limitation.....	5
1.7 Research Methodology	6
1.8 Ethical Consideration	7
1.9 Thesis Organisation.....	8
Chapter 2	10
Background	10
2.1 Introduction	10
2.2 Online Recruitment system	10
2.3 The Importance of Online Recruitment.....	12
2.3.1 Time reduction	13
2.3.2 Cost Reduction	14
2.3.3 Reach huge masses (employers and candidates).....	14
2.3.4 Filter Functions	15
2.3.5 Development of Brand image.....	15
2.3.6 Sophisticated Management Tools.....	16
2.4 E-Recruitment Methods	16
2.5 Online Recruitment Fraud	17
2.6 Knowledge Discovery from Data (KDD).....	18
2.7 Data mining.....	21
2.7.1 Web Mining.....	23
2.7.2 Text Mining.....	26
2.8 Data mining Techniques	28
2.8.1 Classification.....	28
2.8.2 Clustering.....	32

2.8.3	Regression.....	32
2.9	Data mining Tools	33
2.9.1	Rapid Miner (RM)	33
2.9.2	WEKA.....	34
2.9.3	Orange.....	35
2.10	Summary	35
Chapter 3	36
3.1	Introduction	36
3.2	Online Recruitment Related Works.....	36
3.3	Data mining techniques Related works	38
3.4	Summary	49
Chapter 4	50
The Proposed Online Recruitment Fraud Detection Model	50
4.1	Introduction.....	50
4.2	Dataset.....	52
4.3	Pre-processing.....	55
4.4	Selection Feature.....	55
4.5	Classification Algorithm.....	56
4.6	Evaluation Parameter.....	61
4.7	Summary.....	62
Chapter 5.....	63
Results and Discussion.....	63
5.1.	Introduction.....	63
5.2.	Pre-processing Data set.....	63
5.3.	Selection Features.....	66
5.4.	Ensemble Classification.....	68
5.5.	Summary.....	74
Chapter 6.....	75
Conclusion and Future Work.....	75
6.1.	Conclusion.....	75
6.2.	Future Work.....	76
8	Appendix:	86

List of Figures

Figure 2.1. Internet usage in Saudi Arabia until Jan 2018 (Kemp, 2018).....	13
Figure 2.2. KDD Process (Cios, Pedrycz, Swiniarski, & Kurgan, 2007)	18
Figure 2.3. The conventional KDD model (Priyadharsini.C & Thanamani, 2014)	19
Figure 2.4. Classification phases (Gorade, Deo, & Purohit, 2017).....	29
Figure 4.1.Support Vector Machine SVM (DivyaTomar & SonaliAgarwal, 2015).	56
Figure 4.2. Ensemble Classifier - random forest classifier general concept (Securonix Labs,2017) ..	57
Figure 4.3. Random Forest Flow chart	59
Figure 5.1.Best filter Attribute extraction	68
Figure 5.2.Classification accuracy summary	69
Figure 5.3. Confusion matrix for random forest calssification	69
Figure 5.4. Precision, Recall, F-measure and ROC values	70
Figure 5.5. ROC Curve for ensemble classifier.....	71

List of Tables

Table 2.1. Classification Models Comparison	30
Table 4.1. Work Flow of the Proposed Model	51
Table 4.2. Dataset content description	53
Table 5.1. Screen of the one items of the original data set before pre-processing(EMSCAD, 2018) ...	64
Table 5.2. Processed dataset screen.....	66
Table 5.3. Feature extracted via SVM approach	67
Table 5.4. Cross-validation impacts on the classifier performance	72
Table 5.5. Attributes Rank (cross validation 5).....	72
Table 5.6. Attributes Rank (cross validation 10).....	73
Table 5.7. Result Comparison of Current research and Vidros et al. research	74

List of Acronyms

Abbreviations	Explanations
AI	Artificial Intelligence
AUC	Area Under the Curve
CART	Classification and Regression Tree
CIA	Confidentiality Integrity Availability
CV	Curriculum Vitae
EMSCAD	Employment Scam Aegean Dataset
FP	False Positive
IE	Information Extraction
IG	Information Gain
IR	Information Retrieval
KDD	Knowledge Discovery in Database
KNN	K Nearest Neighbour
MLP	Multilayer Perceptron
NB	Naïve Bayes
ORF	Online Recruitment Fraud
RM	Rapid Miner
ROC	Receiver Operating Characteristics
R2L	Remote to Local
ROC	Receiver Operating Characteristics
SVM	Support Vector Machine
U2R	User to Root
WEKA	Waikato Environment for Knowledge Analysis