

Table of Contents

Chapter 1. Introduction	4
1.1. Spam Emails.....	5
1.2. Evolution of Spam Emails.....	7
1.2.1. The Early Year (Manual Spamming).....	7
1.2.2. The Second Phase (Machines for Spamming).....	7
1.2.3. Third Phase (Machine against Machine)	8
1.3. Types of Spam	8
1.3.1. Advertisement Spam.....	9
1.3.2. Financial Spam.....	14
1.3.3. Phishing.....	16
1.3.4. Image Spam	17
1.4. Spam Consequences.....	17
1.5. Spam Filters.....	19
1.6. Spam Filtering Methods	19
1.6.1. Non Technical Solutions....	20
1.6.1.1. Recipient Revolts	20
1.6.1.2. Customer Revolts.....	21
1.6.1.3. Vigilante Attack.....	21
1.6.1.4. Hiding the Addresses.....	21
1.6.1.5. Legitimate Contacts and Limiting Trial Accounts.....	22
1.6.2. Technical Solutions.....	22
1.6.2.1. Domain Filters.....	23
1.6.2.2. Black Listing.....	23
1.6.2.1. White Listing.....	23
1.6.2.2. Rule Based Methods.....	24
1.7. Motivation	25
1.8. Research Objective.....	26
Chapter 2. Literature Survey.....	29
2.1. Manual Inspection	29
2.2. System Approaches	30
2.2.1. Grey Listing.....	31
2.2.2. White Listing.....	31
2.2.3. Black Listing.....	31
2.2.4. Collaborative approach.....	31
2.2.5. Challenge Responce.....	32
2.3. Content based Methods	33
2.3.1. Ad-hoc Rule Based Approach	33
2.3.2. Bayesian Filtering	34
2.4. Machine Learning Research.....	34
2.4.1. Bayesian Classifiers.....	34
2.4.2. Perceptron.....	37
2.4.3. Support Vector Machine.....	39
2.4.4. Nearest Neighbors.....	43
2.4.5. Decision Tree.....	45
2.4.5.1. Alternating Decision Tree.....	46
2.4.5.2. Decision Stump.....	46

2.4.5.3.	Reduced Error Pruning.....	47
2.4.6.	Random Forest.....	47
2.5.	Approaches to Create a Better Classifier	50
2.5.1.	Boosting Algorithms.....	51
2.5.1.1.	Bagging.....	51
2.5.1.2.	Boosting with Re-Sampling.....	52
2.5.1.1.	Adaptive Boosting.....	53
2.5.2.	Combining Classifiers.....	54
2.5.3.	Evolutionary Algorithms.....	56
2.5.3.1.	Genetic Algorithm based classifiers.....	57
2.5.3.2.	Genetic Programming based classifiers.....	57
2.5.4.	Research on Different Part of Spam	62
2.6.	Current Anti-Spam Systems.....	62
2.6.1.	Government Initiatives for Anti-Spam	63
2.6.2.	Industry Oriented Anti-Spam Associations	64
Chapter 3.	Experimental Design	67
3.1.	Structure of Spam Filter	67
3.2.	Different Types of Attacks on Email	69
3.2.1.	Tokenisation.....	69
3.2.2.	Obfuscation	69
3.2.3.	Weak Statistical	69
3.2.4.	Strong Statistical	69
3.3.	Corpora.....	69
3.3.1.	Most Complex Enron Corpus	70
3.3.2.	SpamAssassin	70
3.3.3.	LingSpam.....	71
3.3.4.	Training with Enron (5, 6) and Testing with Enron (All Version)	71
3.4.	Pre-processing	72
3.4.1.	Feature Extraction.....	72
3.4.2.	Dimensionality Reduction	73
3.4.3.	Feature Selection Process	73
3.4.3.1.	Document Frequency.....	73
3.4.3.2.	Information Gain.....	74
3.4.3.3.	Gain Ratio.....	74
3.4.3.4.	Chi-Square.....	75
3.4.3.5.	Relief F.....	75
3.4.3.6.	One Rule.....	76
3.4.4.	Feature Subset Search	76
3.4.4.1.	Genetic Search.....	77
3.4.4.2.	Greedy Stepwise search.....	78
3.4.4.3.	Best First Search.....	78
3.4.5.4.	Rank Search.....	79
3.4.5.	Re-parameterisation	79
3.4.5.1.	Latent Semantic Indexing.....	79
3.5.	Feature Representation.....	80
3.6.	Evaluation Parameters.....	81
3.7.	Discussion	83
Chapter 4.	Feature Selection and Subset Search Methods.....	84
4.1.	Section 1: Evaluation of Best Feature Selection Technique	84
4.1.1.	Aim of This Study.....	84

4.1.2.	Corpora For This Study	85
4.1.3.	Feature Selection Techniques	85
4.1.4.	Classifier for This Study	85
4.1.5.	System Design	85
4.1.6.	Evaluation Metrics.....	85
4.1.7.	Results and Analysis	85
4.2.	Section 2: Feature Subset Search	96
4.2.1.	Aim of This Study.....	96
4.2.2.	Corpora For This Study	96
4.2.3.	Feature Subset Search Techniques.....	96
4.2.4.	Classifiers for This Study.....	96
4.2.5.	System Design	97
4.2.6.	Evaluation Metrics.....	97
4.2.7.	Results and Analysis	97
4.3.	Discussion	100
Chapter 5.	Machine Learning Classifier	102
5.1.	Evaluation of Best Machine Learning.....	102
5.1.1.	Aim of This Study.....	102
5.1.2.	Corpora For This Study	102
5.1.3.	Feature Subset Search Techniques.....	102
5.1.4.	Classifiers for This Study.....	103
5.1.5.	System Design	103
5.1.6.	Evaluation Metrics.....	103
5.1.7.	Results and Analysis	103
5.2.	Discussion	106
Chapter 6.	Machine Learning with Excellent Features.....	108
6.1.	Best Combination of Machine Learning and Features Subset Selection.....	108
6.1.1.	Aim of This Study.....	108
6.1.2.	Corpora For This Study	108
6.1.3.	Feature Subset Search Techniques.....	108
6.1.4.	Classifiers for This Study.....	109
6.1.5.	System Design	109
6.1.6.	Evaluation Metrics.....	109
6.1.7.	Results and Analysis	109
6.2.	Discussion	115
Chapter 7.	Combining and Ensemble Based Classifiers.....	116
7.1.	Section 1: Combining Classifiers with Committee Selection.....	116
7.1.1.	Aim of This Study.....	117
7.1.2.	Corpora For This Study	117
7.1.3.	Feature Subset Search Techniques.....	117
7.1.4.	Classifiers for This Study.....	117
7.1.5.	System Design	117
7.1.6.	Evaluation Metrics.....	118
7.1.7.	Results and Analysis	118
7.2.	Discussion	131
7.3.	Section 2: Enhanced Genetic Programming Classifier	133
7.3.1.	Aim of This Study.....	134
7.3.2.	Corpora For This Study	134
7.3.3.	Feature Subset Search Techniques.....	135
7.3.4.	Classifiers for This Study.....	135

7.3.5.	System Design	135
7.3.6.	Evaluation Metrics.....	135
7.3.7.	Results and Analysis	136
7.4.	Discussion	140
Chapter 8.	Training, and Testing Time	141
8.1.	Evaluation of Robust Email Filtering Models.....	141
8.1.1.	Aim of This Study.....	141
8.1.2.	Corpora For This Study	142
8.1.3.	Feature Subset Search Techniques.....	142
8.1.4.	Classifiers for This Study.....	142
8.1.5.	System Design	142
8.1.6.	Evaluation Metrics.....	142
8.1.7.	Results and Analysis	143
3.4.	Discussion	151
Chapter 9.	Conclusion, Business Implication and Limitation	152
9.1.	Business Implications.....	153
9.2.	Contribution of this Research.....	154
9.3.	Future Work	155
9.4.	Limitations	155
References	157
Appendix A:	Features selected by Re-Parametrization Methods	170
Appendix B:	Features selected by Feature Selection Method (Relief F).....	172
Appendix C:	Features selected by Feature Subset Search Method (Greedy Stepwise).....	174