SPAM DETECTION IN EMAIL BODY USING HYBRID OF ARTIFICIAL NEURAL NETWORK AND EVOLUTIONARY ALGORITHMS

ALI OTMAN ALI ALBSHAYREH

A project report submitted in partial fulfillment of
the requirements for the award of the degree of
Master of Computer Science (Information Security)

Faculty of Computing
Universiti Teknologi Malaysia

JANUARY 2015

To my beloved Father and Mother who have great merit on my, for their unceasing love and relentless patience and devotion raising me.

To my lovely wife that she was so patience and for giving me her love, tender, caring, support, confidence, and hope.

To my roses, daughters, who are whole my life

To all my brothers and sisters who have stood by me.

To my respected supervisor, Dr. Maheyzah Md Siraj

To my beloved homeland, Jordan.

Thank you…!

ACKNOWLEDGEMENT

First and foremost, I give thanks and praise to Allah for his direction and blessings and for granting me knowledge, fortitude, and determination in the successful achievement of this research work and project.

My heartfelt thanks and earnest gratitude are due to Dr. Maheyzah Md Siraj for her constant support and guidance and for their enduring patience.

I would like also to express my appreciation to all the people at Computer Science Faculty at UTM who contributed to my success one way or the other.

My deepest love and supreme appreciation and gratitude to my parents for their unceasing love and relentless patience and devotion raising me.

To my wife that she was so patient and for giving me her love, tender and caring. My thanks go to my brothers and sisters as well, for their love and encouragement.

Thanks for to all my friends and whom I cannot thank enough for their continual encouragement and unvarying support and assistance, and for their true friendship and loyalty through the thick and thin.

Finally, to all the people who have had a positive impact in my life:

Thank you…!

ABSTRACT

Spam detection is a significant problem that is considered by many researchers through various developed strategies. Creating a particular model to categorize the wide range of spam categories is complex; with understanding of spam types, which are always changing. In spam detection, low accuracy and the high false positive are substantial problems. So the trend to hire a global optimization algorithm is an appropriate way to resolve these problems due to its ability to create new solutions and non-compliance with local solutions. In this study, a hybrid machine learning approach inspired by Artificial Neural Network (ANN) and Differential Evolution (DE) are designed for effectively detect the spams. Comparisons have been done between ANN-DE with Genetic Algorithm (GA) and ANN-DE with InfoGain algorithm to show which approach has the best performance in spam detection. Spambase dataset of 4061 E-mail in which 1813 were spam (39.40%) and 2788 were non-spam (59.60%) were used to training and testing on these algorithms. The popular performance measure is a classification accuracy, which deals with false positive, false negative, accuracy, precision, and recall. These metrics were used for performance evaluation on the hybrid of ANN-DE with GA and InfoGain algorithm as feature selection algorithms. Performance of ANN-DE with GA and ANN-DE with InfoGain are compared. The experimental results show that the proposed hybrid technique of ANN-DE and GA gives better result with 93.81% accuracy compared to ANN-DE and InfoGain with 93.28% accuracy. The results recommend that the effectiveness of proposed ANN-DE with GA is promising and this study provided a new method to practically train ANN for spam detection.

ABSTRAK

Pengesanan spam adalah masalah penting yang dianggap oleh ramai penyelidik melalui pelbagai strategi maju. Mewujudkan model tertentu untuk mengkategorikan pelbagai jenis kategori spam adalah kompleks; dengan memahami jenis spam, yang sentiasa berubah-ubah. Dalam pengesanan spam, ketepatan yang rendah dan positif palsu yang tinggi adalah masalah besar. Jadi, trend untuk mendapatkan algoritma pengoptimuman global adalah cara yang sesuai untuk menyelesaikan masalah-masalah ini disebabkan keupayaannya mencipta penyelesaian baru dan tidak mematuhi penyelesaian tempatan. Dalam kajian ini, pendekatan pembelajaran mesin hibrid diilhamkan oleh Rangkaian Neural Buatan (ANN) dan Berbeza Evolution (DE) direka dengan berkesan untuk mengesan spam. Perbandingan dilakukan antara ANN-DE dengan Algoritma Genetik (GA) dan ANN-DE dengan algoritma InfoGain bagi memastikan pendekatan mana mempunyai prestasi terbaik dalam pengesanan spam. Dataset Spambase yang mempunyai 4061 emel di mana 1813 adalah spam (39.40%) dan 2788 adalah bukan spam (59,60%) telah digunakan untuk latihan dan ujian ke atas algoritma ini. Ukuran prestasi popular adalah ketepatan klasifikasi, terdiri daripada positif palsu, negatif palsu, ketepatan dan keingatan. Metrik tersebut digunakan untuk penilaian prestasi hibrid ANN-DE dengan GA dan InfoGain sebagai algoritma pemilihan ciri. Prestasi ANN-DE dengan GA dan ANN-DE dengan InfoGain dibandingkan. Keputusan eksperimen menunjukkan bahawa teknik hibrid ANN-DE dan GA yang dicadangkan memberikan hasil yang lebih baik dengan ketepatan 93.81% berbanding ANN-DE dan InfoGain dengan ketepatan 93.28%. Keputusan ini mencadangkan bahawa keberkesanan ANN-DE dengan GA yang dicadangkan adalah membanggakan dan kajian ini menyediakan kaedah baru yang praktikal untuk melatih ANN untuk pengesanan spam.

TABLE OF CONTENTS

CHAPTER	TITLE	PAGE
	DECLARATION	ii
	DEDICATION	iii
	ACKNOWLEDGMENT	iv
	ABSTRACT	v
	ABSTRAK	vi
	TABLE OF CONTENTS	vii
	LIST OF TABLES	xi
	LIST OF FIGURES	xii
	LIST OF ABBREVIATION	xiv
1	**INTRODUCTION**	
	1.1 Overview	1
	1.2 Background of project	3
	1.3 Problem statement	5
	1.4 Aim of project	5
	1.5 Objectives of the project	6
	1.6 Scope of the project	6
	1.7 Significant of the project	7
	1.8 Organization of the project	7
2	**LITERATURE REVIEWS**	
	2.1 Introduction	8
	2.2 Spam History Overview	8
	2.3 Spam Definition and General Characteristics	9

2.4 Type of Spam 12

2.5 Issues in Spam Detection 14

2.6 Solution Methodologies 15

2.6.1 Anti-Spam Legislation Efforts 15

2.6.2 E-Mail Transmission Protocols 16

2.6.3 Local Changes in Email Transmission Process 17

2.6.4 Spam Filtering 17

2.6.4.1 Structure of a Learning Based Spam Filters 19

2.7 Feature Selection Techniques in Spam Detection 22

2.7.1 Genetic Algorithm (GA) 26

2.8 Existing Spam Filter Technique 28

2.8.1 Non-Machine Learning 29

2.8.1.1 Whitelist 29

2.8.1.2 Blacklistt 30

2.8.1.3 Greylist 31

2.8.1.4 Traffic Analysis 31

2.8.1.5 Collaborative Spam Filtering 32

2.8.2 Machine Learning 32

2.8.2.1 Unsupervised Machine Learning 33

2.8.2.2 Supervised Machine Learning 36

2.8.3 Artificial Neural Network (ANN) 37

2.8.3.1 Description of ANN and the way it Works 38

2.8.3.2 Learning Paradigms in (ANN) 43

2.8.3.3 Related Works 45

2.9 Deferential Evolution (DE) 48

	2.10	Justification on Hybrid ANN with DE and using GA	50
	2.11	Summary	51
3	**RESEARCH METHODOLOGY**		
	3.1	Introduction	52
	3.2	Research Framework	52
	3.3	Phases of Project	54
		3.3.1 Phase I: Normalization and features Selection	55
		3.3.2 Phase II: Implement of Improved ANN-DE	56
		3.3.2.1 Differential Evolution Discretion	56
	3.4	Phase III: Evaluate the performance of (ANN-DE)	60
	3.5	Dataset discretion	62
	3.6	Summary	63
4	**FEATURE SELECTION**		
	4.1	Introduction	64
	4.2	Features Selection	64
		4.2.1 InfoGain Feature Selection Algorithm	65
		4.2.2 Genetic Algorithms in Feature Selection (GA)	66
		4.2.2.1 Fitness Function	69
	4.3	Classification with ANN	69
	4.4	Selected Features	73
	4.5	Summary	80
5	**IMPLEMENTATION AND RESULTS**		
	5.1	Introduction	81
	5.2	The examination overview	81
	5.3	Implementation of ANN	82

	5.4	Implementation of Hybrid ANN-DE	85
	5.5	Discussion on Results	91
	5.6	Summary	97
6	**CONCLUSION AND RECOMMENDATIONS**		
	6.1	Introduction	98
	6.2	Project Achievement and limitations	99
	6.3	Future Work and Recommendations	100
	6.4	Summary	100
	REFERENCES		**102**

LIST OF TABLES

TABLE NO.	TITLE	PAGE
2.1	Different Spam Definitions	10
2.2	Categories of Spam	13
2.3	The strong and drawback of feature synthesis and feature subset selection	24
2.4	The different between wrapper and filter supervised feature selection	25
2.5	Related Works	45
3.1	Table of DE Parameters	60
3.2	False Positive and False Negative	61
3.3	Performance Measures with a Definition and Formula	61
3.4	Spambase dataset characterization	62
3.5	Sample of the spambase dataset	63
4.1	The parameters values used in ANN	71
4.2	ANN testing results with diverse parameter values	72
4.3	Testing results of ANN classifier using InfoGain attribute evaluation	74
4.4	Genetic Algorithm parameters of experimental environment	76
4.5	The most significant selected features using GA and there accuracy in different population size	77
5.1	Testing result of ANN via Subset-1 and Subset-2	84
5.2	Testing result of ANN-DE via Subset-1 and Subset-2	87
5.3	Compare the proposed technique with previous works	96

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE
2.1	Technical Anti-Spam Approaches	18
2.2	The Architecture Of Spam Filter	19
2.3	E-mail message construction from perspective feature extraction	20
2.4	Description of pre-processing steps	21
2.5	Some of popular feature selection approaches	23
2.6	GA Crossover operation	27
2.7	GA Mutation operation	27
2.8	Various approaches to spam filtering	29
2.9	The structure of the Fuzzy ART ANN	36
2.10	The structure of natural neurons	38
2.11	The artificial neuron structure	39
2.12	The structure of ANN.	40
2.13	Some of the activation functions used in ANN	42
2.14(a)	ANN with feed forward	42
2.14(b)	ANN competition with feedback	42
2.15	Common structure of the DE algorithm	49
3.1	Research Framework	53
3.2	Research workflow	54
3.3	Pseudo code of (DE)	59
4.1	Simple K-NN example	67
4.2	The GA-KNN feature weighting	68
4.3	ANN accuracy rate with different parameters value	73
4.4	Testing results of ANN classifier and using InfoGain attribute evaluation	75

4.5	Accuracy of GA in attribute evaluation	80
5.1	The examination overview procedure	82
5.2	ANN flowchart	83
5.3	ANN implementation procedure	84
5.4	ANN-DE implementation procedure	86
5.5	Classification accuracy of ANN and ANN-DE with different spambase subsets	88
5.6	Classification FP of ANN and ANN-DE with different spambase subsets	88
5.7	Classification FN of ANN and ANN-DE with different spambase subsets	89
5.8	Accuracy of ANN using Subset-1and Subset-2	91
5.9	False Positive of ANN using Subset-1and Subset-2	92
5.10	False Negative of ANN using Subset-1and Subset-2	92
5.11	Accuracy of ANN and ANN-DE using Subset-1	93
5.12	False Positive of ANN and ANN-DE using Subset-1	93
5.13	False Negative of ANN and ANN-DE using Subset-1	94
5.14	Accuracy of ANN, GA-ANN and ANN-DE	95
5.15	False Positive of ANN, GA-ANN and ANN-DE	95
5.16	False Negative of ANN, GA-ANN and ANN-DE	96

LIST OF ABBREVIATIONS

A	–	Accuracy
AIS	–	Artificial Immune System
ANN	–	Artificial Neural Network
ANN-DE	–	Hybrid Artificial Neural Network with Differential Evolution algorithm
ARMM	–	Automated Retroactive Minimal Moderation
ART	–	Adaptive Resonance Theory
BoW	–	Bag-of-Words
BP	–	Back Propagation
DE	–	Differential Evolution
DK	–	Domain Keys
DKIM	–	Domain Keys Identified Mail
DMP	–	Designated Mailers Protocol
DNSBL	–	Domain Name System Black-Hole List
DT	–	Decision Tree
FFNN	–	Feed Forward Neural Network
FN	–	False Negative
FP	–	False Positive
FTC	–	Federal Trade Commission
GA	–	Genetic Algorithm
GSA	–	Gravitational Search Algorithm
GSA	–	Gravitational Search Algorithm
IIM	–	Identified Internet Mail
InfoGain	–	Information Gain
ISP	–	Internet Serves Provider
KNN	–	K Nearest Neighbor Algorithm

MDA	–	Mail Delivery Agent
MLP	–	Multilayer Perceptron
MSE	–	Mean Square Error
MSTP	–	Simple Mail Transfer Protocol
MTA	–	Mail Transfer Agent
NB	–	Naïve Bayes
P	–	Precision
PSO	–	Particle Swarm Optimization
R	–	Recall
SA	–	Simulated Annealing
SOFM	–	Self-Organizing Feature Map
SOM	–	Self-Organizing Map
SPF	–	Sender Policy Framework
SPF	–	Sender Permitted From
SVM	–	Support Vector Machine
TEOS	–	Trusted E-mail Open Standard
TN	–	True Negative
TP	–	True Positive
TREC	–	Text Retrieval Conference
UBE	–	Unsolicited Bulk Email
UCE	–	Unsolicited Commercial Email

CHAPTER 1

INTRODUCTION

1.1 Overview

As a global network, internet refers to computers that work much like the postal system at sub-second speeds. Just as if the postal service allows people to exchange envelopes containing messages among themselves, E-mail is one of the services provided by the Internet, which enables users to communicate despite distances via a free and convenient medium. As we all know, there is a no full secure system because the presence of intrusive and vandals. A phenomenon of spam is one of the most problems that threat the feasibility of communication via email (Kågström, 2005).

In fact, there is no precise and specific definition of spam, but all combines on it is a bad thing just to hear this term. (Merriam-Webster Online Dictionary, 1994) explained spam as "uninvited usually commercial e-mail directed to a large number of addresses". Accordingly, Spam is most often consider to be electronic junk mail or junk newsgroup postings (Almeida *et al.*, 2010). To some people, spam can be considered more generally as any unsolicited email. Real spam is commonly email advertising, sent to a mailing list or newsgroup. In most cases that have been monitored the aim of spam can be politically or religiously, fool the recipients with promises of wealth, or have viruses may harm the receiver computer. However, if you are signed-up one of the websites, it is immediately send an email to your registered email to confirmation on the registration. In most cases are dealt with this e-mail as spam, even though if one

can talk over that what is unsolicited mail for one user can be an exciting email message for another.

Spam email has been increasingly problem, which cause for users exhaust a time to take out the messages from inbox, while it ,may be unintentional deleted of significant emails may cause a financial losses and even able to cause limit the mailbox space and wasting network resources. Based on what has been published by McAfee in March 2009, while the spam filter working at 95% accuracy , the cost of lost in productivity for the user almost $0.50 per day, based on user having to expend 30 seconds per day to deal with two spam messages. Consequently, around $182.50 was the annual losses of productivity per employee (Almeida*, et al.*, 2010)

According to recent study by Kaspersky Lab as a IT Security company, more than 70.7% of emails sent in second quarter (April - June) 2013 were really spam, an increase of more than 4.2% over first quarter (January - March) 2013 totals. Spam-O-Meter is free source of spam statistical and live tools to quantify the actual amount of spam in the internet profess 87.4% of all e-mail messages are spam up to this time compared to 86.6% a year ago. The incremental number of spam in email, many methods have been suggested to filter spam email (Allias *et al.*, 2014; Gudkova, 2013). Ferris Research expected the worldwide cost of spam to be in 2005 and 2007 US$ 50 billion and US$ 100 billion respectively (Bauer *et al.*, 2008).

Because of these reasons and for the outright infringement on personal space, some of preventing the normal delivery of spam is required. The gravity of the spam problem can be inferred from the fact that. On December 16, 2003 the U.S. House of Representatives had ratified the revised version of bill, as part of the CAN-SPAM Act, to restrict unwanted messages by imposition a financial penalties of $ 6 million and five years prison (Lee, 2005a; Sivanadyan, 2003).

Distinguish how spam has developed about whether and break down the development of spam filters. However, that not permitted to take a gander at each kind

of spam filters. By performing this analysis, it may be feasible for spam filters to stay one-stage in front of spammers and conceivably even put an end to spam in the end.

Subsequently, there are numerous associations, and additionally people, who have taken it upon themselves to resist spam with several methods. By and by, the internet is open, and there is little that might be done in order to restrain spam, in the same way that it is unfit to turn away junk-mail. In any case, the steady expand movement of spam needs to be expanded labours to hinder, which has created significant fear for clients.

1.2 Background of Problem

With the proliferation of the Internet and e-mail service, has become essential in our lives and be inherent to most of us. With this deployment, spam has been appeared, which is a growing problem and the constant threat of users and carriers, which drain time, money, and network resources, on the other side there is no cost where identical messages sent to various recipients by spammer.

For the entire algorithms of an email classification, there is the problem of discovery, a sensible interchange amongst two kinds of errors: categorising genuine mail as spam (False Positive) and classifying spam as genuine mail (False Negative). When spam messages are classified as legitimate mail, the user becomes irritated, the reverse situation can lead to the real loss of valuable information. Therefore, in spam detection, low accuracy and the high false positive are substantial issue.

The duty of classification is difficult and continuously altering (Carpinter and Hunt, 2006). Creating a particular model to categorize the wide range of spam categories is complex; with understanding of spam types, which are always changing made the task of classification near impracticable. In addition, the majority of users find out that false positives and low accuracy are not acceptable. The ongoing

evolution of spam could be partly credited to alerting tastes and trends in the market; though, spammers frequently change their mails to keep away from detection techniques, adding up additional obstruction to true detection.

The efforts had been started and increased to innovates and develops techniques to classifying the emails whether they are legitimate or spam. It is fit has spammers to periodically alternate their skills, approach, actions, and to fake their messages, to keep away from these techniques. Various spam classification techniques had been offered based on email contents. for instants, Bayesian analysis (Sahami *et al.*, 1998), heuristic approaches (Cook *et al.*, 2006), machine learning approaches (Guzella and Caminhas, 2009). Therefore, the task requires rapidity response to the constant evolution of the spammer by take advantage of user feedback to be employed in the learning algorithm (Bratko *et al.*, 2006).

Machine learning is one of overall techniques proposed for spam detection and had achieved success and breathtaking results. Unfortunately, in machine learning, a high dimensionality of characteristics space in the wake of preprocessing have to be as an enormous obstruction for the classifier. Add to this, the intemperate number of characteristics additionally can debase the classification; this was due to large amount of words in the messages that can extracted (Allias*, et al.*, 2014).

The discovery of the back-propagation technique, a lot of modified and new algorithms have been proposed for training feed-forward neural networks; many of these algorithms having a very fast convergence rate for reasonable size networks. However, there was a dramatic decrease in the number of appropriate network training algorithms when the neural network training becomes a large scale, i.e. the number of network parameters develops considerably. For example, learning a large number of hidden layer weights in a multi-layer perceptron (MLP) neural network can be regarded as a large-scale optimization problem. Conversely, global optimization methods are under continuous development, and recently, studies have identified genetic algorithms and evolution strategies as promising stochastic optimization methods.

1.3 Problem Statement

A significant issue concerning spam classifiers is the trouble of error convergence, i.e. the decrease in error amongst the desired and calculated unit values. Many issues to be considered in classification process, precisely the Artificial Neural Network (ANN), such as big data size that may contains irrelevant and redundant features , the ANN design, the training algorithm, ANN parameters training, initial weights, etc., which decrease the error rate. All these will influence the convergence of ANN learning and may be trapped the ANN in a local minima solution. So the trend to employ global optimization algorithms in an appropriate way to resolve these problems as in preprocessing phase as feature selection algorithms (Combinations improvement) and during the classification process as trainer algorithm (Collaborative improvement).

The main objectives and desired outcome in this project are to accessed following:

i) How to pre-process and prepare the email features?

ii) How to optimize a classifier that can recognize spam email?

iii) How to evaluate and validate the performance of the classifier?

1.4 Aim of the Project

This project seeks to improve detection accuracy and reduce the false positive rate in spam detection. This will be achieved by implementing hybrid Artificial Neural Network (ANN) with Differential Evolution (DE), seeking to efficiently (ANN) train parameter.

1.5 Objectives of the Project

This study will focus on increasing the accuracy and reducing the false positive of the spam detection techniques based on classification of email content. The scope of the project is as the following:

i) To select significant features from dataset by applying InfoGain algorithm and Genetic Algorithm (GA).
ii) To develop a classifier to detect spam email based on Genetic Algorithm (GA) as a features evaluator and hybrid of Artificial Neural Network (ANN) and Differential Evolution (DE) as a trainer algorithm.
iii) To evaluate the proposed classifier (ANN-DE) and (ANN) classifier using InfoGain algorithm and Genetic Algorithm (GA) as features selection, depending on the accuracy, false positive, false negative.

1.6 Scope of the Project

This study will focus on increasing the accuracy and reducing the false positive of the spam detection techniques based on classification of email content. The scope of the project is as the following:

i) The implementation of Artificial Neural Network (ANN) and the hybrid of ANN and Differential Evolution (DE) will be done based on the email content and using InfoGain algorithm and Genetic Algorithm (GA) as a feature selection techniques.
ii) Spambase datasets will be used; it will be collected from UCI website (http://archive.ics.uci.edu/ml/datasets/).
iii) The performance will be evaluated based on accuracy, false positive, false negative, recall and precision.

1.7 Significant of the Project

The amount of spam emails has amplified. Therefore, users need to prevent their email. Different method has been created to impact spam, for the automatically spam detection technique which are responsible to eliminate such spam email from a user email box. Take into consideration spam trouble, users must own filtration technique, which can serve them to categorize the email into spam or otherwise. Spam can intrude a user mailbox with no agreement of the user. By applying the projected technique in this study which is the hybrid of ANN and DE, users will gain a valuable techniques and help them to maximise the amount of blocked spam to their mailbox.

1.8 Organization of the Project

This project is organized into four chapters; first chapter contains the introduction, problem background, statement of problem, aim of the project, project objectives, scope of the project and significant of the study. The second chapter describes literature on spam such as spam definition, type of spam also; it includes a review of spam detection technique. The third chapter explains the methodology of the project that will be used to achieve objectives. The initial result will be discussed on the fourth chapter with a brief analysis. The fourth chapter discusses the process of implementation. Chapter five covers the statistical analyses of results. Lastly, chapter six summarizes the whole study.

REFERENCES

Alander, J. T. 1994. *An indexed bibliography of genetic algorithms: Years 1957-1993*: Art of CAD.

Allias, N., Megat, M. N., Noor, M. and Ismail, M. N. 2014. A hybrid Gini PSO-SVM feature selection based on Taguchi method: an evaluation on email filtering. Paper presented at the *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication*, Siem Reap, Cambodia.

Almeida, T. A., Yamakami, A. and Almeida, J. 2010. Probabilistic anti-spam filtering with dimensionality reduction. Paper presented at the *Proceedings of the 2010 ACM Symposium on Applied Computing*, Sierre, Switzerland.

Alpaydin, E. Introduction to Machine Learning. 2004. *Cover, Copyright Page, Table of Contents for*, 1-327.

Androutsopoulos, I., Koutsias, J., Chandrinos, K. V. and Spyropoulos, C. D. 2000. An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. Proceedings of the 2000 *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 160-167.

Arram, A., Mousa, H. and Zainal, A. 2013. Spam detection using hybrid Artificial Neural Network and Genetic algorithm. Proceedings of the 2013 *Intelligent Systems Design and Applications (ISDA), 2013 13th International Conference on*. 8-10 Dec. 2013. 336-340.

Basavaraju, M. and Prabhakar, D. R. 2010. A novel method of spam mail detection using text based clustering approach. *International Journal of Computer Applications*, 5(4), 15-25.

Bauer, J. M., Van Eeten, M. J. and Chattopadhyay, T. 2008. ITU Study on the Financial Aspects of Network Security: Malware and Spam. *ICT Applications*

and Cybersecurity Division, International Telecommunication Union, Final Report, July.

Blanzieri, E. and Bryl, A. 2008. A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review*, 29(1), 63-92. doi: 10.1007/s10462-009-9109-6

Bratko, A., Filipič, B., Cormack, G. V., Lynam, T. R. and Zupan, B. 2006. Spam filtering using statistical data compression models. *The Journal of Machine Learning Research*, 7, 2673-2698.

CAPTCHA. 2005. The CAPTCHA project. http://www.captcha.net,.

Carpenter, G. A., Grossberg, S. and Rosen, D. B. 1991. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural networks*, 4(6), 759-771.

Carpinter, J. and Hunt, R. 2006. Tightening the net: A review of current and next generation spam filtering tools. *Computers & security*, 25(8), 566-578.

Catalin, C. 2006. An anti-spam filter based on adaptive neural networks. Proceedings of the 2006,

Chhabra, S. 2005. *Fighting spam, phishing and email fraud.* Master Thesis, UNIVERSITY OF CALIFORNIA.

Christina, V., Karpagavalli, S. and Suganya, G. 2010. A Study on Email Spam Filtering Techniques. *International Journal of Computer Applications*, 12(1), 0975-8887.

Cook, D., Hartnett, J., Manderson, K. and Scanlan, J. 2006. Catching spam before it arrives: domain specific dynamic blacklists. Proceedings of the 2006 *Proceedings of the 2006 Australasian workshops on Grid computing and e-research-Volume 54*, 193-202.

Cormack, G. V. 2007. Email spam filtering: A systematic review. *Foundations and Trends in Information Retrieval*, 1(4), 335-455.

Cormack, G. V. and Lynam, T. R. 2005. TREC 2005 Spam Track Overview. Proceedings of the 2005 *TREC*,

Corne, D., Dorigo, M., Glover, F., Dasgupta, D., Moscato, P., Poli, R., et al. 1999. *New ideas in optimization*: McGraw-Hill Ltd., UK.

Cranor, L. F. and LaMacchia, B. A. 1998. Spam! *Communications of the ACM*, 41(8), 74-83.

Dhanaraj, S. and Karthikeyani, V. 2013. A study on e-mail image spam filtering techniques. Proceedings of the 2013 *Pattern Recognition, Informatics and Medical Engineering (PRIME), 2013 International Conference on*, 49-55.

Drake, C. E., Oliver, J. J. and Koontz, E. J. 2004. Anatomy of a Phishing Email. Proceedings of the 2004 *CEAS*,

Duan, Z., Dong, Y. and Gopalan, K. 2005. A differentiated message delivery architecture to control spam. Proceedings of the 2005 *Parallel and Distributed Systems, 2005. Proceedings. 11th International Conference on*, 255-259.

Emanuelsson, O., Nielsen, H. and Heijne, G. V. 1999. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Science*, 8(5), 978-984. doi: 10.1110/ps.8.5.978

Freeman, J. A. 1994. *Simulating neural networks with mathematica*. Reading, Mass.: Addison-Wesley.

Garro, B. A., Sossa, H. and Vázquez, R. A. 2010. Design of artificial neural networks using differential evolution algorithm *Neural Information Processing. Models and Applications* (pp. 201-208): Springer.

Gentleman, R. and Carey, V. J. 2008. Unsupervised Machine Learning *Bioconductor Case Studies* (pp. 137-157): Springer New York.

Gomes, L. H., Cazita, C., Almeida, J. M., Virg, #237, Almeida, l., et al. 2004. Characterizing a spam traffic. Paper presented at the *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, Taormina, Sicily, Italy.

Goodman, J., Cormack, G. V. and Heckerman, D. 2007. Spam and the ongoing battle for the inbox. *Communications of the ACM*, 50(2), 24-33.

Gudkova, D. 2013. Spam in Q2 2013. Available from: http://www.securelist.com/en/analysis/204792297/Spam_in_Q2_2013.

Gulyás, C. 2006. *Creation of a Bayesian network-based meta spam filter, using the analysis of different spam filters.* Master's Thesis, Citeseer.

Guzella, T. S. and Caminhas, W. M. 2009. A review of machine learning approaches to spam filtering. *Expert Systems with Applications*, 36(7), 10206-10222.

Holland, J. H. 1975. *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*: U Michigan Press.

Hulten, G., Penta, A., Seshadrinathan, G. and Mishra, M. 2004. Trends in Spam Products and Methods. Proceedings of the 2004 *CEAS*,

Hunt, R. and Carpinter, J. 2006. Current and new developments in spam filtering. Proceedings of the 2006 *14th IEEE international conference on networks, ICON*, 1-6.

Hurink, J. 2001. Introduction to Local Search. *Lecture notes*.

Idris, I. and Selamat, A. 2014. Improved email spam detection model with negative selection algorithm and particle swarm optimization. *Appl. Soft Comput.*, 22, 11-27. doi: 10.1016/j.asoc.2014.05.002

Ilonen, J., Kamarainen, J.-K. and Lampinen, J. 2003. Differential evolution training algorithm for feed-forward neural networks. *Neural Processing Letters*, 17(1), 93-105.

Isa, D., Kallimani, V. and Lee, L. H. 2009. Using the self organizing map for clustering of text documents. *Expert Systems with Applications*, 36(5), 9584-9591.

John, G. H., Kohavi, R. and Pfleger, K. 1994. Irrelevant Features and the Subset Selection Problem. Proceedings of the 1994 *ICML*, 121-129.

Kågström, J. 2005. *Improving naive bayesian spam filtering.* Master thesis, Mid Sweden University.

Kohavi, R. and Sommerfield, D. 1995. Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology. Proceedings of the 1995 *KDD*, 192-197.

Kuipers, B. J., Liu, A. X., Gautam, A. and Gouda, M. G. 2005. Zmail: zero-sum free market control of spam. Proceedings of the 2005 *Distributed Computing Systems Workshops, 2005. 25th IEEE International Conference on*, 20-26.

Lai, C.-C. and Tsai, M.-C. 2004. An empirical performance comparison of machine learning methods for spam e-mail categorization. Proceedings of the 2004 *Hybrid Intelligent Systems, 2004. HIS'04. Fourth International Conference on*, 44-48.

Lai, G.-H., Chen, C.-M., Laih, C.-S. and Chen, T. 2009. A collaborative anti-spam system. *Expert Systems with Applications*, 36(3), 6645-6653.

Langton, C. G. 1997. *Artificial life: An overview*: Mit Press.

Lee, Y. 2005a. The CAN-SPAM Act: a silver bullet solution? *Communications of the ACM*, 48(6), 131-132.

Lee, Y. 2005b. The CAN-SPAM Act: a silver bullet solution? *Commun. ACM*, 48(6), 131-132. doi: 10.1145/1064830.1064863

Levi, M. and Wall, D. S. 2004. Technologies, Security, and Privacy in the Post-9/11 European Information Society. *Journal of law and society*, 31(2), 194-220.

Levine, J. R. 2005. Experiences with Greylisting. Proceedings of the 2005 *CEAS*,

Li, K., Pu, C. and Ahamad, M. 2004. Resisting SPAM Delivery by TCP Damping. Proceedings of the 2004 *CEAS*,

Liang, J., Yang, S. and Winstanley, A. 2008. Invariant optimal feature selection: A distance discriminant and feature ranking based solution. *Pattern Recognition*, 41(5), 1429-1439.

Lugaresi, N. 2004. European Union vs. Spam: A Legal Response. Proceedings of the 2004 *CEAS*,

Luger, G. F. 2005. *Artificial intelligence: Structures and strategies for complex problem solving*: Pearson education.

Luo, X. and Zincir-Heywood, N. 2005. Comparison of a SOM based sequence analysis system and naive Bayesian classifier for spam filtering. Proceedings of the 2005 *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*, 2571-2576.

Masters, T. and Land, W. 1997. A new training algorithm for the general regression neural network. Proceedings of the 1997 *Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on*, 1990-1994.

Mezura-Montes, E., Velázquez-Reyes, J. and Coello Coello, C. A. 2006. A comparative study of differential evolution variants for global optimization. Proceedings of the 2006 *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, 485-492.

Mitchell, T. M. 1997. *Machine learning. New York: McGraw-Hill Education*.

Nagamalai, D., Dhinakaran, C. and Lee, J. K. 2007. Multi layer Approach to defend DDoS attacks caused by Spam. Proceedings of the 2007 *Multimedia and Ubiquitous Engineering, 2007. MUE'07. International Conference on*, 97-102.

Pedersen, M. E. H. 2010. Good parameters for differential evolution. *Magnus Erik Hvass Pedersen*.

Prechelt, L. P. 1994. A Set of Neural Network Benchmark Problems and Benchmarking Rules.

Priddy, K. L. and Keller, P. E. 2005. *Artificial neural networks: an introduction* (Vol. 68): SPIE Press.

Punch III, W. F., Goodman, E. D., Pei, M., Chia-Shun, L., Hovland, P. D. and Enbody, R. J. 1993. Further Research on Feature Selection and Classification Using Genetic Algorithms. Proceedings of the 1993 *ICGA*, 557-564.

Riedel, J. 2004. *The Evolution of Spam and SpamAssassin A Major Qualifying Project Report submitted to the Faculty of the.* WORCESTER POLYTECHNIC INSTITUTE.

Rogalsky, T., Kocabiyik, S. and Derksen, R. 2000. Differential evolution in aerodynamic optimization. *Canadian Aeronautics and Space Journal*, 46(4), 183-190.

Saad, P. 2003. Enhancement of neural network convergence with hidden layer and memory part control. In. *Proceeding of the International Conference on Robotics, 2003, Pulau Pinang.*

Sahami, M., Dumais, S., Heckerman, D. and Horvitz, E. 1998. A Bayesian approach to filtering junk e-mail. Proceedings of the 1998 *Learning for Text Categorization: Papers from the 1998 workshop, 1998. Madison, Wisconsin: AAAI Technical Report WS-98-05, 98-105.*, 98-105.

Saito, T. 2005. Anti-spam system: another way of preventing spam. Proceedings of the 2005 *Database and Expert Systems Applications, 2005. Proceedings. Sixteenth International Workshop on*, 57-61.

Salehi, S. and Selamat, A. 2011. Hybrid simple artificial immune system (SAIS) and particle swarm optimization (PSO) for spam detection. Proceedings of the 2011 *Software Engineering (MySEC), 2011 5th Malaysian Conference in*. 13-14 Dec. 2011. 124-129.

Sarangi, P. P., Sahu, A. and Panda, M. 2013. A Hybrid Differential Evolution and Back-Propagation Algorithm for Feedforward Neural Network Training. *parity*, 84(14).

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1), 1-47.

Siedlecki, W. and Sklansky, J. 1990. A note on genetic algorithms for large-scale feature selection. *Pattern recognition letters*, 10(5), 335-347.

Sivanadyan, T. 2003. Spam? not any more! detecting spam emails using neural networks: Technical report, University of Wisconsin.

Stumberger, G., Dolinar, D., Palmer, U. and Hameyer, K. 2000. Optimization of radial active magnetic bearings using the finite element technique and the differential evolution algorithm. *Magnetics, IEEE Transactions on*, 36(4), 1009-1013.

Subramaniam, T., Jalab, H. A. and Taqa, A. Y. 2010. Overview of textual anti-spam filtering techniques. *International Journal of the Physical Sciences*, 5(12), 1869-1882.

Templeton, B. 2003. Origin of the term "spam" to mean net abuse. *Essays on Junk E-mail (Spam)*.

Tetko, I. V., Livingstone, D. J. and Luik, A. I. 1995. Neural network studies. 1. Comparison of overfitting and overtraining. *Journal of Chemical Information and Computer Sciences*, 35(5), 826-833.

Tretyakov, K. 2004. Machine learning techniques in spam filtering. Proceedings of the 2004 *Data Mining Problem-oriented Seminar, MTAT*, 60-79.

Twining, D., Williamson, M. M., Mowbray, M. and Rahmouni, M. 2004. Email Prioritization: Reducing Delays on Legitimate Mail Caused by Junk Mail. Proceedings of the 2004 *USENIX Annual Technical Conference, General Track*, 45-58.

Unler, A. and Murat, A. 2010. A discrete particle swarm optimization method for feature selection in binary classification problems. *European Journal of Operational Research*, 206(3), 528-539.

Unler, A., Murat, A. and Chinnam, R. B. 2011. mr^2 PSO: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. *Information Sciences*, 181(20), 4625-4641.

Vapnik, V., Golowich, S. E. and Smola, A. 1997. Support vector method for function approximation, regression estimation, and signal processing. *Advances in neural information processing systems*, 281-287.

Wang, X.-h. and He, Y.-g. 2005. A new neural network based power system harmonics analysis algorithm with high accuracy. *Power System Technology*, 29(3), 72-75.

Widrow, B. and Lehr, M. A. 1990. 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. *Proceedings of the IEEE*, 78(9), 1415-1442.

Wu, C.-H. 2009. Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks. *Expert Systems with Applications*, 36(3), 4321-4330.

Yegnanarayana, B. 2009. *Artificial neural networks*: PHI Learning Pvt. Ltd.