Chapter 13
Malay Language Text-Based Anti-Spam System Using Neural Network

Hamid A. Jalab
University Malaya, Malaysia

Thamarai Subramaniam
University Malaya, Malaysia

Alaa Y. Taqa
Mosul University, Iraq

ABSTRACT
This unauthorized intrusion has cost time and money for businesses and users. The exponential growth of spam emails in recent years has resulted in the necessity for more accurate and efficient spam filtering. This chapter focuses on creating a text-based anti-spam system using back-propagation neural network for Malay Language emails that efficiently and effectively counter measure spam problems. The proposed algorithm consists of three stages; pre-processing, implementation and evaluation. Malay language emails are collected and divided into spam and non-spam. Features are extracted and document frequency as dimension reduction technique is calculated too. Classifiers are trained to recognize spam and non-spam emails using training datasets. After training, classifiers are tested to check whether they can predict spam (or non-spam) emails accurately with the testing datasets. The result of this classification in terms of accuracy, precision, and recall are evaluated, compared and analyzed, thus providing the best anti-spam solution to counter measure spam problem of Malay language emails.

INTRODUCTION
Email is one of the most popular communication tools that were ever invented. It has proliferated internet usage since it was introduced, and allows users to communicate with each other at low cost while providing an efficient message delivery system. However, the simplicity and low cost in sending an email has paved the way for unsolicited emails. Individual users and businesses can send
thousands of emails to recipients at any given time. These emails, also known as spam, are unsolicited emails, which are neither requested nor required by the recipients. Spam either contains harmless marketing information or malicious codes such as viruses that could cause data loss, thus leading to inconvenience and/or economic loss to the recipients. Unsolicited emails are widely viewed as a serious threat to the internet as it clogs up the users’ inboxes and cost businesses billions of dollars in wasted bandwidth (Cournane & Hunt, 2004). To combat spam, researchers and developers have created many anti-spam tools. The basic function of anti-spam tools is to filter emails by separating spam from genuine mail and adding them into a junk mail box. Various methods and standards are used to fight spam nuisances (Subramaniam, Jalab & et al, 2010).

There were many studies carried out on spam filtering that were effective and efficient on detecting and blocking spam email. However, these studies were mainly performed on English language email spam. Methods (preprocessing and Machine learning algorithms) used for English language spam detection will limit the performance of a classifier given the nature of different human languages (Özgür, Güngör & et al, 2004; Pang, Feng, & et al 2007). (Özgür et al., 2004) proposed dynamic spam filtering methods based on Artificial Neural Network and Bayesian algorithms for agglutinative language and for Turkish in particular (which is a complex morphology). They performed five different experiments by using Single Layer Perceptron (SLP), Multi-Layer Perceptron (MLP) and Bayesian with 3 different feature vector sizes. Their experiments showed that some non-Turkish words that occurred frequently in spam mail were better classified than most Turkish words.

(Dong, Cao & et al, 2006) indicated that segmenting Chinese words (email) restricts the performance of existing spam filter. They used Bayesian spam filter based on cross N-gram on CCERT Computer Emergency Response Term of which 940 were spam emails and 1400 were non-spam Chinese language emails. These emails were then partitioned into 10 parts. 5 characters of crossed N-gram and three different feature selection methods were used: Mutual Information, Odd Ratio and $X^2$–statistic ($CHI$). Comparison of all 3 feature selection methods were reported based on $\text{»} = 1$ and $\text{»} = 9$. They concluded that the Odds Ratio selection scheme produced the best result and errors can be further reduced with the combination of rule-based methods.

(Pang & et al., 2007) used Support Vector Machine by adopting the tri-gram language model for word segmentation of Chinese emails and applied Discount Smoothing algorithms to overcome the sparse data problem. Automaton Machine identifies different factoid words. They experimented using LingSpam (English email) and CCERT data sets of Chinese emails, and made comparison between Maximum Entropy, Bayesian, Bayesian with Good-Turning, Bayesian with Absolute Smooth and Support Vector Machine.

(Anh, Anh & et al, 2008) specified that token segmentation of the Bayesian filter produced less effective performances for detecting Vietnamese language-based spam. Therefore, they proposed a Vietnamese segmentation for token selection based on language classification and Bayesian. They implemented two filters; token segmentation based on whitespaces and token selection based on the Vietnamese segmentation approach. The result showed that Vietnamese segmentation token selection coupled with Bayesian classifier generated more effective spam detection - 9% more accurate as compared to other segmentation techniques.

(Na Songkhla & Piromsopa, 2010) proposed a statistical rule-base for Thai language spam detection for Spam Assassin that can be shared among servers and easily maintained. They used Chulalongkorn University’s Thai Word segmentation program that separated Thai words more accurately with low processing time and conditional probability (CP) to find spam-like patterns. They concluded that Spam Assassin’s,