Design and Development of Decision Making Model for Spam email Classification Using Neural Network

Ms. Dipalee Patil
Department of Computer Science & Engineering
MSS's College Of Engineering & Technology,
Jalna.
dipalee.patil78@gmail.com

Prof. Anil Turukmane
Department of Computer science & Engineering
MSS's College of Engineering & Technology,
Jalna.
anilturukmane@gmail.com

Abstract:- Internet has changed the way of communication, which has become more and more concentrated on emails. Emails, text messages and online messenger chatting have become part and parcel of our lives. Out of all these communications, emails are more prone to exploitation. Thus, various email providers employ algorithms to filter emails based on spam and ham. In the proposed system, the prime aim is to detect text as well as image based spam emails. To achieve the objective, Artificial Neural Network is applied for the classification of spam and ham emails. Preprocessing of email text before executing the algorithms is used to make them predict better. The system uses Enron corpus's dataset of spam and ham emails. In this system, the performance of ANN will be measured based on four measuring factors namely: precision, sensitivity, specificity and accuracy.

Keywords- *Spam, Ham, KNN, Nai've Bayes, Artificial Neural Network, Image Spam.*

*****

I. Introduction

As number of internet users is increasing day by day, morepeople are finding email communication an inexpensive way tosend their data and communicate with their peers. With pros also come some cons. This is evident from thefact that spam emails have accounted for 68.8% of all emailtraffic in 2012.The increasing numbers of spam emails not only wastes one's time but also wastes network resources significantly.

With the increasing importance of the email and the intrusions of internet marketers, unsolicited commercial email (spam) has become a major problem on the internet [13]. Spam has caused serious economy loss and become a social issue. Unwanted email can come from anywhere. Current trends say that 95 percent of Internet transmission will be Spam [14]. This type of image spam accounts for 40% of all global spam in 2007 compared with just 1% in late 2005.

Spam arises from an online social situation and nowadays it becomes a social problem. Although current anti-spam technologies are quite successful in filtering text based spam emails[17]. A new trend in email spam is the emergence of image spam. The image spam is substantially more difficult to detect, as they employ a variety of image creation and randomization algorithms .In Image spam the text message is embedded into attached images to defeat the anti spam filters.

There are two main types of spam and they have different affects on Internet users. Cancellable Usenet spam is a single message sent to 20 or more Usenet groups. Usenet spams aims at "lurkers", people who read newsgroups but rarely or never post and give their address away. Usenet spam subverts the ability of system administrator to manage the topics they accept on their systems. Another type of Email spam targets individual users with direct mail messages. Email spam is any email that meets the following three criteria:

- Anonymity: The address and identity of the sender are concealed.
- Mass Mailing: The email is sent to large group of people.
- Unsolicited: The email is not requested by recipients.

Spam Mail has become an increasing problem in recent years. It has been estimated that around 70% of all emails are spam.

These unsolicited bulk emails are coined by the term "Spam". Spam email with advertisement text embedded in images generally known as image spam, which poses a great challenge to Anti spam filters in detecting these spam emails [17]. E-mail management has become a vital and growing problem for individuals and organizations as it is prone to misuse

In the past, spam filtering required the manual construction of pattern matching rule sets [18].Bayesian classifiers to learn spam characteristics. OCR-based modules can be used against image spam, to tolerate the analysis of the semantic content embedded into images. The main limitation of this

OCR-based spam classification technique is that it requires more processing time.

II. Literature Survey

The use of internet has been extensivelyincreasing over the past decade and it continues to be on the ascent..

2.1 Content-based ImageSpam Filtering

According to the latest report of MAA WG in 2011 [20],about 90% emails were spam all over the world,To detect spam based on the textualcontent of the email, many text-based anti-spam approacheshave been proposed, such as Bayesian filters and SupportVector Machine (SVM) filters. In order to solve the problem of image spam, G .Fumeraetal. [21] first introduced an approach that analyzes the textinformation embedded in images with the help of Optical Character Recognition (OCR) technique. Typically, content obscuring techniques such asCAPTCHA were used to invalidate OCR tools [22].

2.2 Machine Learning Applied ForEmail Deception

Sujeet More and Dr S A Kulkarni focused mainly on cognitive (spam) words for classification This method, can be easily implemented, compares amiably with respect to popular algorithms, like Logistic Regression, Neural Network, Naive Bayes and Random Forest using polynomial kernel as filter.In recent years, due to increasing use of e-mail whichhas led to the emerging and further escalation of problemscaused by deceptive e-mail messages, commonly referred to as spam mail.

2.3 Classifying Spam Emails usingText and Readability Features

Rushdi Shams and Robert E. Mercer has proposed a method which utilizes text features that are long established such as frequency of spam words and HTML tags as well as some that are new. In addition, features related to message readability (e.g., reading difficulty indexes, complex and simple word frequency, document length, and word length) are used. The features are extracted from four standard email datasets—CSDMC2010, SpamAssassin, LingSpam, and Enron-Spam.Another notable finding is that the classifier induced by BAGGING performs the best on all of the datasets.

2.4 Active Learning Feedback-Driven Semi-Supervised Support Vector Machine

Jian Zhong, YiLu Zhou, Wei Deng has proposed a method which is inspired by natures of image-based spams:large quantity, similarity and character variability. Spammers have to send their em ails in large quantity from same IPs orIP Sub net. If the mail has been identified as spam by the rules it willbeblocked. Otherwise, if it is image-based mail, it willbecaptured by the plug-in.

2.5 Supervised Learning using Machine Learning Techniques

Ms.D.KarthikaRenuka ,Dr.T.Hamsapriya , Mr. M. Raja Chakkaravarthi and Ms.P. Lakshmi Surya have proposed a different machine learning techniques for spam classification**.** From address, to address, type of spam received, organization from which the spam was received were few of the attributes used.For analyzing real time dataset and to predict the performance, the supervised learning algorithms were adopted here [1]. There are two main paradigms for handling an ensemble of different classification algorithms: Classifier Selection and Classifier Fusion.

2.6 Classification of Image Spam Using Artificial Neural Networks

M. Soranamageswari, Dr. C. Mena, has presented an experimental system for the classification of image spam by considering statistical image feature histogramand mean value of an block of image. The authors have concentrated more on measures of statistical feature i.e. color histogram and mean More specifically a feed forward neural network, namely a multilayer perceptron istrained to predict directly JPEG, GIF, PNG, Bit map images.

2.7 Open source OCR system

There has been a resurgence of interest in optical character recognition (OCR) in recent years, driven by a number of factors. Search engines have raised the expectation of universal access to information on line, and cheap networking and storage have made it technically and economically feasible to scan and store the books, newspapers, journals, and other printed materials of the world. OCRopus is a new, open source OCR system emphasizing modularity, easy extensibility, and reuse, aimed at both the research community and large scale commercial document conversions.

III. MOTIVATION

Global spam volume increased very fast over the past fiveyears. E-mail spam accounted for 96.5% of incoming e-mailsreceived in business by June 2008 .Nucleus researchreports that spam e-mail, on average, costs U.S organizations$874 per person annually in lost productivity. The success of text-based spam filtering techniques has driven spammers to find new variations of spam, and their latest invention is the image spam. As reported by McAfee [23], image spam accountsfor approximately 30% of all e-mail spam.

The main objectives of the spammers are:

- To generate profit in the form of money: spammerssend advertisements embedded in the spam e-mails.
- To promote products and services: companies dealwith the spammers and pay them to promote theirproducts and services.

- To steal sensitive information such as credit card numbers, passwords and bank account details: thismay be achieved in two ways (i) back door entry created by malicious programs and(ii) launching phishing attacks.

E-mail spam is a growing and serious problem faced by e-mailadministrators and users. Early image spam simply embedded advertising text in images that linked to HTML formatted e-mail. To deal with image spam, filtering technologies began to incorporate Optical Character Recognition (OCR)into the filters to detect the text in the images.

IV. Problem Definition

Designing of a software system which can detect the text as well as image based spam and protect user and computer from the damages like extensive consumption of bandwidth, overload of mail server and wastage of time in detecting spam mails by implementing a decision making model using Artificial Neural Network.

V. Objectives

The main objectives of the proposed system are:

- To detect text as well as image based spam: The proposed system should detect both, the text as well as image based spam.
- To achieve more accuracy over the preprocessing phase:The preprocessing phase is the most important phase in the proposed system. If the data is preprocessed properly then only it can be used to infer accurately. The ambiguous data may generate the inefficient output, so it should be avoided.
- To train the neural network with maximum training data set to make itmore efficient: The Enron corpus dataset will be used in the proposed system. The proposed system should be trained with the variety of spam as well as non spam mails. It should be trained for maximum number of text and image based spam mails.

VI. Proposed Model

The proposed system will have the following three phases:

- Black Listing and White listing
- Pre-processing
- Classification

6.1 Black Listing and White listing

Black Listing: Black-listing is creating a list of domainnames which are used by the spammers, when a mail comesfrom that specific domain which is black listed it is considered spam. No further processing is done.

White Listing: White list is a list of trusted domains and a mailfrom them is always ham. White listing is a method used toclassify user's email addresses as legitimate ones.its listing is not always accurate.Therefore, to counter all these techniques employed by spam filters, spammers now send mails with embedded imagescontaining the spam text.

6.2 Extracting words from Image

To extract the text out of these images is an arduous task. It must be done by sophisticated OCR toolsand based on the high level, low level. .All those web pages and domains that are notorious for sendingspam mails and are not trusted; go on the list of black list . Further, spam isin the eye of the recipient, so a white list is maintained where users can mark those websites they want mails from whetherthey send "spam" or not. Optimum accuracy is achieved for a clear resolution image and more popular fonts like Times New Roman.

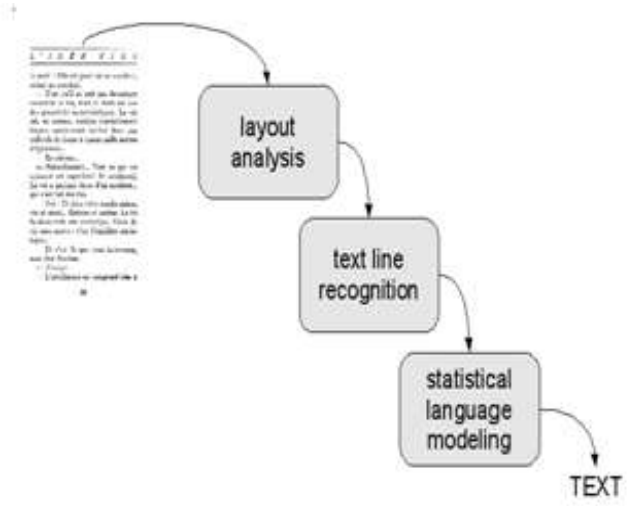

Figure 6.1 : Flow diagram of OCR engine

The overall architecture of the OCR system is a strictly feed-forward architecture (no backtracking) with three major components: (physical) layout analysis, text line recognition, and statistical language modeling;

- Physical layout analysis is responsible for identifying text columns, text blocks, text lines, and reading order.
- Text line recognition is responsible for recognizing the text contained within each line (note that lines can be vertical or right-to-left) and representing possible recognition alternatives as a hypothesis graph.
- Statistical language modeling integrates alternative recognition hypotheses with prior knowledge about language, vocabulary, grammar, and the domain of

the document. Text line recognition itself either relies on black-box text line recognizers

6.2.1 Dataset

A large set of email messages, the Enron corpus, was made public during the legal investigation concerning the Enron Corporation. Enron corpus containsa total of 200,399 messages belonging to 158 users with an average of 757 messages per user. This is approximately one third the size of the original corpus.

6.3 Porter Stemmer algorithm

A database of all the words that occur in each mail with the frequency of the word stored in each column will be maintained. So it is converted to their root form first by applying Porter Stemmer algorithm.

Some steps of this algorithm are:

- Remove the plurals and -ed or –ing suffixes
- Deal with suffixes , -full, -ness etc.
- Take off -ant, -ence etc.

After the database with the stemmed words,with each mail name in one column and the frequency of occurrence of words in other the system will move on to the next phase.

6.4 Neural Network

A neural network is a set of connected input/output units in which each connection has a weight associated with it.Back propagation is a neural network learning algorithm.The Back propagation algorithm performs learning on a multilayer feed-forward neural network. During The learning phase, the network learns by adjusting the weights so as to be able to predict the correct class lable of the input tuples.

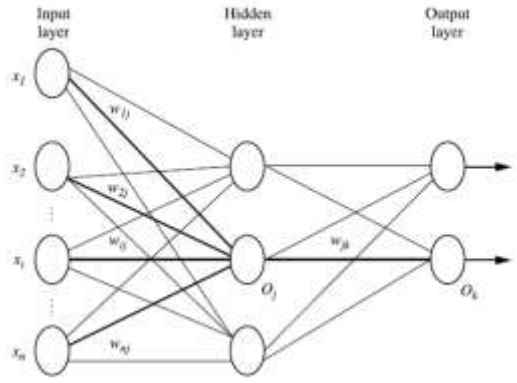

Figure 6.1: Feed forward Neural Network

Each layer is made up of units. The inputs to the network correspond to the attributes measured for each training tuple. The inputs are fed simultaneously into the units making up the input layer. These inputs pass through the input layer and are then weighted and fed simultaneously to a second layer of "neuron like" units, known as a hidden layer.

Similarly, a network containing two hidden layers is called a three-layer neural network, and so on. The network is feed-forward in that none of the weights cycles back to an input unit or to an output unit of a previous layer. It is fully connected in that each unit provides input to each unit in the next forward layer. Each output unit takes, as input, a weighted sum of the outputs from units in the previous layer.

References

[1] Harisinghney A. ; Dixit A. ; Gupta S. ; Arora A. "Text and Image based spam email classification using KNN, naïve Bayes and Reverse DBSCAN algorithm" Optimization, Reliability and Information Technology(ICROIT) , 2014 International Conference on DOI:10.1109/ICROIT.2014.6798302, page(s):153-155, 2014

[2] N. Nhung and T. Phuong."An Efficient Method for Filtering Image-Based Spam E-mail". Proc. IEEE International Conference on Research, Innovation and Vision for the Future (RIVF07), IEEE Press, Mar. 2007 , pp. 96-102. doi: 10.1I 0 9 /RIVF.2007.369I 4 1.

[3] Ketari, Lamia Mohammed, Munesh Chandra, and Mohammadi Akheela Khanum. "A Study of Image Spam Filtering Techniques."Computational Intelligence and Communication Networks (CICN), 2012 Fourth International Conference on IEEE,2012.

[4] R. Smith, "An Overview of the Tesseract OCR Engine," in Proc. International Conference on Doc ument Analysis and Recognition, 2007

[5] Klimt, Bryan, and Yiming Yang. "The Enron corpus: A new dataset for email classification research." Machine learning: ECML 2004. Springer Berlin Heidelberg, 2004. 217-226.