Machine learning algorithm to identifies fraud emails with feature selection

Anita Sindar Sinaga¹*, Musthafa Haris Munandar², Arjon Samuel Sitio³

¹Informatics Engineering, STMIK Pelita Nusantara, Indonesia
²Faculty of Science and Technology, Labuhanbatu University, Indonesia
³Informatics Engineering, STMIK Pelita Nusantara, Indonesia

*haito_ita@yahoo.com

Abstract One percent of the emails that come in each day are fraudulent. Promotion emails tend to offer products. The recipient's email is recorded by a company or organization. Not all promotional emails are considered spam or hoaxes. When observed, incoming promotional emails provide the information that is needed. How to identify promotional emails including hoaxes or shipping with machine learning, known as algorithms, Support Vector Machines, Naïve Bayes, Decision Tree, Logistic Regression, Stochastic Gradient Descent, and Neural Network (MLP). Decision trees are effective tracing using a data structure consisting of vertices & edges. A node (root, branch, leaf) can categorize incoming e-mail, including hoax or e-mail shipping. Previously, it was necessary to categorize the characteristics of the email. After searching the email, the promotional email grouping is continued. In Feature Extraction, the calculation of Gain and Entropy is used to determine the selection of features in the classification of promotional emails, fraudulent emails or hoaxes.

1. Introduction

Google email account has provided spam or shipping email features. Email users can group emails as desired. Spam email is sent to the inbox without inviting the sender. Email is sent to more than one person. The promotional email feature is business, the recipient's account is recorded after registering or subscribing to the company or organization [1]. Spam that goes to inbox accounts is Phishing Email and Hoax [2]. Promotion e-mails contain the necessary information but there are also shipping e-mails, it looks as if the actual e-mail contains important information after being traced is a hoax. The incoming email statistics show that 36% of all sent spam email contains advertisements. Types of email spam include promotional emails, phishing emails, gift offering emails, hoax emails and fraudulent emails [3]. The fraudulent email contains document attachments such as pdf waiting for the bait from the e-mail recipient to open and download the malware [4]. Most email content consists of blackmail by offering the lowest prices, lucrative sales and promises in catchy language without being suspicious [5].

Reinforcement Learning is part of machine learning algorithms requiring agents to act in the environment with the decision tree concept converting data into decision rules [6]. The decision tree is developed from the assumption that attributes provide information on the relationship between classes and attribute values [7] [8]. Grouping starts from selecting attributes for the root tree and branches for attribute values. The process stops when the leaf is not branched, there is the same instance.
The proposed solutions classify emails, positive or negative content, analyze the happiness of email recipients, use several machine learning algorithms for spam detection and subject-based email classification [9]. This study tested 600 phishing emails and 400 harmless emails using Natural Language Processing Techniques, one by one the letters were processed resulting in an email classification as phish or nonphish [10].

2. Methodology

Emails that enter the promotional inbox are first grouped by email subject, it can also be grouped from the sender of the email. If you tend to promote without being related to the information needed, the email response is spam. Furthermore, the incoming email processing scheme is to categorize spam or not.

![Figure 1. Steps of Promotions Email Classification](image)

A method commonly used for machine learning data processing:

a. Multilayer Perceptron (MLP)
Artificial Neural Networks compiled with the structure and function of the human brain as a model to emulate. Multilayer Perceptron is a topology, the most common of Artificial Neural Networks, connected perceptrons to form several layers (layers). An MLP has input layer, at least one hidden layer and output layer.

b. Naïve Bayesian classifier
The Naive Bayes Classifier is a simple probability classification process which refers to the Bayes Theory which states that the likelihood of an event occurring is equal to the intrinsic probability (calculated from currently available data) times the probability that something similar will based on knowledge that happened in the past.

c. Decision tree classifier
The result of grouping from decision tree C4.5 is structural data in the form of binary trees 1 and 0. Modeling the initial data set of C4.5 tree is a set of basic tuples to determine the classes associated with these tuples. It is assumed that the tuples belong to a pre-grouped class categorized by the adjective labeled in the class label. then the formation of training sets is carried out randomly taken from the base. This technique is most efficient using classification extensively. The tree structure is formed from:

a) To make decision tree i.e. must calculate root nodes, internal nodes and leaf nodes. This root node will be retrieved from the attribute has the highest gain value, by calculating the entropy value at the whole case and calculate the value entropy and gain in each attribute.

b) These internal nodes will be retrieved from attributes which has the highest gain value by removing the attributes it has selected as the root node. This leaf node will take from the attribute with the gain value highest attribute value only entered into one condition.

The use of feature selection is carried out in preprocessing for data modeling features. In the data analysis stage, there are two groups in the application of feature selection, namely ranking selection and subset selection. the subset selection method is used to find a set of features that are considered as optimal features. The subset selection method is used for selection by wrapper type (outer), selection by filter type and selection with embedded type (inner). Use of rank in grouping to determine the
ranking level independently between selected features. The highest features will be used and the lowest features will be ignored.

3. Result and Discussion
The fraudulent e-mail identification process as follows:

a. Hoax: IP Address Server, one-way, provocative title, URL address, facts, authenticity of photos.

b. Scams email: in the form of documents, gifts, certain registrations, fake domains, payments, extortion.

Distinguishing fraudulent e-mail in promotional e-mails is done by processing the data set, namely the incoming e-mail in the Gmail inbox. Selecting fraudulent emails with the C45 decision tree begins by categorizing the types of email as fraudulent. filter is done from the identity of the sender, email header, email recipient. Usually fraudulent emails contain colorful writing containing attractive offers and promising prizes. Use of filter headers by extracting information from email headers. Each filter backlist will determine spam messages and stop all emails coming from the backlist file. It can be seen that rule-based filters will more strongly recognize senders via subject line using user defined criteria. on the granting filter sends messages with the recipient's pre-approval. Finally, Challenge-response filter performed by applying an algorithm for getting the permission from the sender to send the mail.

![Figure 2. Process of Email Filtering System](image)

The method used in selecting training data and data testing is to use the K-fold cross validation. determine training data and testing data from the whole data. K-fold cross validation repeats k-times to divide a random data set into k-subsets are independent of each other, each replication is left with one a subset for testing and another subset for training. The recommended value for K is 10 because it is more accurate in its estimated size.

| Table 1. Examples of Training Data That Has Passed The Data Preprocessing |
|---------------------|-----------------|-----------------|--------|-----------------|
| **Sender**         | **Content**       | **Header**       | **Fact** | **Receiver**   | **Payment** |
| Samsung Electronics Indonesia | on Sale            | the things you    | Yes     | No              | Yes         |
| ComputerWeekly.com                          | Read the buyer's guide | why making sense of data makes good business sense | No      | Yes             | No          |
| Data + AI Summit Europe Team                      | Data Team’s Unit    | jules’ picks for data + ai summit europe | No      | Yes             | No          |

The flow chart in the decision tree structure, each internal node describes the attributes of the training data, each branch describes the results of the training data and the leaf node shows the class or class distribution. Number of incoming e-mails = 325, advertising data 93, Education data 177, Platform data 55. of the equation \( k = 1 + 3,322 \log n \), The value 8 becomes the number of classes or labels for the attribute Advertisement. Data on total email ads per month is at most = 228 minus the total data for email ads per month at least = 107, then the interval \( c \) email ads in iteration one (S1) is 15 iterasi. Parameter weights, calculation of the score to be used to determine the interval for fraudulent emails (in the form of categories) obtained from the multiplication of the label with the weight for each parameter applicable to all training data in each iteration. Score = Label \( \times \) Bobot Parameter. The score interval is used in determining fraudulent email: \( 1,00\leq x \leq 5,5 \) not fraudulent email ; \( 5,6 \leq x \leq 9,5 \) fraudulent email.
Tabel 2. Percentage of Fraudulent Emails

| Criteria         | Parameters | Average % | Final Weight |
|------------------|------------|-----------|--------------|
| Content          | x1         | 8         | 10%          |
| Header           | x2         | 33        | 41%          |
| Receiver         | x3         | 27.5      | 35%          |
| Fact             | x4         | 11        | 14%          |
| Payment Request  | x5         | 0.15      | 6%           |

From the incoming email data, the column taken as an attribute / decision variable is survive, while the columns (attributes) taken by the determination of the decision tree are: email sender, content, one way, fact collect recipient, payment advertisement. Determination of fraudulent e-mail between "Fraud" and "No Fraud":

a. Incoming promotional email contains unnecessary advertisements.
b. The sender of the e-mail is clear but the document contains instructions.
c. Email asking for bio.

Stages of the process of calculating the C45:

1. Calculation of Entropy Value
   The initial step of the C4.5 algorithms is to find the entropy value. First, first determine the total entropy value in the case with the formula:
   \[ \text{Entropy} (S) = \sum_{i=1}^{n} -p_i \times \log_2 p_i \]

2. Calculation of Entropy Value
   The initial step of the C4.5 algorithms is to find the entropy value. First, first determine the total entrepreneurial value in the case with the formula:
   \[ \text{Gain} (S,A) = \text{Entropy} (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \text{Entropy} (S) \]

   The decision tree method converts very large facts into a decision tree that represents the rules. Decision trees can be used to explore data by finding hidden relationships between the number of candidate input variables and a target variable. 1. Prepare training data. Training data is usually taken from historical data that has happened before or is called past data and has been grouped into certain classes. From the incoming email on 1 Gmail account, the column taken as an attribute / decision variable while the columns (attributes) taken by the determination of the decision tree are Content, Header, Receiver, Fact, Payment.

![Figure 3. Decision Tree Iterasi](image)

Based on the last decision tree, we get a rule with the following rules, e-mail that fits the facts is not a hoax or scam email. The sender of fraudulent emails containing advertisements or sales. URL, Domain fake including fraud. Requirements used in choosing a decision tree model:

a. The model that has the most rules and covers the entire test data target class (requirements 1 and 3), namely the 11th iteration model.
b. High accuracy model. Iterations 1-7 and 8-15, the error rate is 0%.
c. A model that includes all target classes may appear in the test set. All rules in iterations 1-15. The selected rule will be applied to the entire data to obtain the level of fraud in promotional emails, namely the payment request attribute.

**Figure 4. Email Promotions Classification**

The choice of features in machine learning with the C45 aims to classify data sets from based training data modeling. The highest feature is used as the basis for data classification while the lowest feature is ignored.

**Figure 5. Steps of Selection Features**

Information Gain in machine learning is used to measure how relevant a feature is on measurement results. The use of this technique can reduce feature dimensions by measuring the Entropy reduction before and after separation.

**Figure 6. Gain and Entropy Feature Selection Results**

4. Conclusion
1. Based on the criteria for grouping fraudulent emails on the promotion email inbox, the header contains 41% fraud, 35% received email from more than one person, 8% requests to make a payment. 14% of promotional emails are true. 90% of fraudulent emails on promotional emails.
2. Gain and Entropy Feature Selection produces hoax and fraud classification. Fill in the no fraud email, then the email is categorized as a hoax.
References

[1] M. Das and V. Prasad, “A Nalysis of an I Mage S Pam in E Mail B Ased on,” vol. 3, no. 3, pp. 129–140, 2014.
[2] M. Sharma, “A Survey of Email Spam Filtering Methods,” vol. 7, pp. 14–21, 2018.
[3] S. Rosi, “Spam and Email Detection in Big Data Platform Using,” vol. 7, no. 4, pp. 53–58, 2018.
[4] H. S. Jones, J. N. Towse, N. Race, and T. Harrison, “Email fraud: The search for psychological predictors of susceptibility,” PLoS One, vol. 14, no. 1, pp. 1–15, 2019.
[5] M. Nishad, “A Review Paper on Phishing Through E-Mail,” no. 1, pp. 1–2, 2018.
[6] A. Bhowmick and S. M. Hazarika, “Machine Learning for E-mail Spam Filtering: Review, Techniques and Trends,” no. June, 2016.
[7] F. Shahbazi, “Using decision tree classification algorithm to design and construct the credit rating model for banking customers,” IOSR J. Bus. Manag., vol. 21, no. 3, pp. 24–28, 2020.
[8] A. K. Shrivas, “Decision Tree Model for Classification of E-mail Data with Feature Selection,” vol. 4840, no. August, pp. 15–19, 2015.
[9] A. Sinha, “Manipulation of Email Data Using Machine Learning and Data Visualization Manipulation of Email Data Using Machine Learning and Data Visualization,” no. October, 2020.
[10] S. Aggarwal, V. Kumar, and S. D. Sudarsan, “Identification and detection of phishing emails using natural language processing techniques,” ACM Int. Conf. Proceeding Ser., vol. 2014-September, no. September 2014, pp. 217–222, 2014.