JuTISI (Jurnal Teknik Informatika dan Sistem Informasi)
Vol 6 No. 1 | April 2020

Fuzzy Logic Approach for Email SPAM Detection System

Damian Prihadi[1], Vivin Trisyanti[2]
[1] damianp.unhas@mailinator.com, [2]vivin.trisyanti@gmail.com

Abstract. The more incessant online business world is accompanied by the increasing number of cases of SPAM email distribution. SPAM Email is the mass distribution of emails with topics and recipients that are not relevant or expected. The emergence of spam really annoys e-mail users because it can increase bandwidth usage, it also significantly reduces storage capacity on the server.
This research was conducted with the aim of utilizing fuzzy logic in the e-mail spam detection process. It is hoped that with this method the clean email selection process from SPAM can be done accurately. The process in this study consisted of 3 parts, namely the initial e-mail selection process, the extraction of prospective clusters, and the formation of clusters. The final result of the whole process is the formation of a cluster or group of e-mails that are identified as SPAM and those that are not SPAM.

***Keywords*- SPAM detection, fuzzy logic, feature selection**

I. INTRODUCTION

The increase in email use is increasing along with the rise of the online business world. Email has replaced the official medium of correspondence between online companies, as well as with their customers. The use of e-mail is also influenced by economic factors (cheap) and speed of delivery to the recipient. But the increase in the use of e-mail has also been accompanied by an increase in the abuse of e-mail to seek unilateral gain.

One of them is the emergence of an unexpected commercial e-mail or what is more commonly known as spam. The emergence of spam is very disturbing to e-mail users because it can increase the usage of internet connection bandwidth, and will become a pile of garbage, thereby reducing storage capacity. Some spam is used to deliver an advertising message that contains pornographic content or a medium for spreading viruses. For that we need a media that can detect and filter spam, so that it can separate spam e-mail.

Research to detect the presence of spam has been developed. Some of them are research on spam campaign detection [1]. Appavu et al. Divides spam email into several campaigns which then give a score according to the criteria of each campaign. Research using fuzzy integrated with Wordnet has been developed [2]. The use of fuzzy is able to provide a solution when an important term that can be a keyword in email and spam detection rarely appears in a collection of e-mail and spam and has a small frequency, so the fuzzy term is generated to be used as a keyword. Apart from using fuzzy clustering logic by using association rule mining [3]. Research [4] also try to detect e-mail and spam using machine learning.

The advantages possessed by association rule mining in e-mail clustering and spam to determine the relationship between terms, as well as the advantages of fuzzy in e-mail and spam clustering by integrating linguistic term variables into a fuzzy set, can be developed by combining fuzzy and association rule mining. The combination of fuzzy and association rule mining [5], namely Fuzzy Frequent Itemset-Based Hierarchical Clustering (F2IHC) is able to increase the level of accuracy and produce overlapping clusters in document clustering.

The use of Adapative Neuro Fuzzy Inference System is also used for weather forecasting [6] [7], the method used combines ANFIS with the cyclic and moving average methods. Based on the advantages possessed by fuzzy logic and association rule mining in detecting and classifying e-mail and spam by determining the relationships that occur between terms through association rules and integrating linguistic term variables into a fuzzy set, the hybrid method between fuzzy logic and association can be developed for rule mining of e-mail and spam detection.

III. METHODS

The system design in this study consists of three main parts, namely: e-mail and spam preprocessing, feature selection, cluster construction with naïve Bayes. An overview of the system architecture is shown in Figure 1.

ISSN 2443-2210 (Print)
ISSN 2443-2229 (Online)

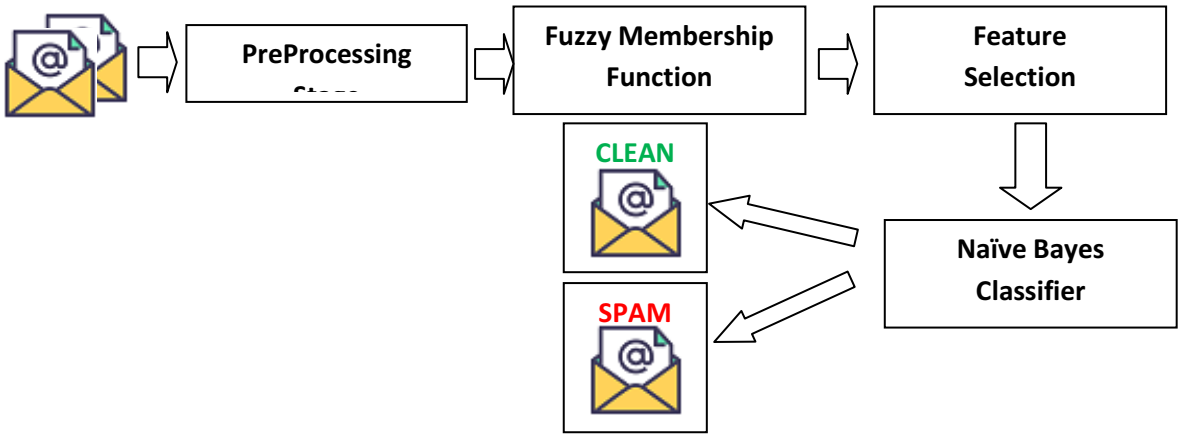

Figure 1. The system architecture

a. Email and Spam Prepocessing

At the preprocessing stage, there is input in the form of a collection of e-mails and spam, a list of stop - word words, minimum tf - idf (α). The stages carried out in preprocessing, namely: term extraction, stop words removal, stemming, and term selection. From the preprocessing results, a collection of keyterms and their frequencies will be obtained

b. Feature Selection

At the feature selection stage, four types of input are used, namely: key terms collection, fuzzy membership function, minimum support, minimum confidence. There are four processes that must be followed to get a candidate cluster, including: calculating the membership function value with a type-2 fuzzy set, finding candidate-1 itemset, finding candidate-2 itemset, and selecting candidate clusters.

c. Cluster Construction with Naïve Bayes

At this stage there are 2 main types of sub-sections, namely: Naïve Bayes Classifier and Classifier Evaluator. At the Naïve Bayes Classifier stage, probabilistic calculations will be carried out based on keywords obtained from feature selection. Meanwhile, the Classifier Evaluator stage is determining the trend of the email whether it is SPAM or non-SPAM.

III. RESULTS AND DISCUSSION

This chapter will explain the results of the trials and evaluations of the methods proposed in this study. The method in this research was applied supported by hardware and software with the specifications of the Intel® Core ™ 2 Duo Processor T5750@2.00Ghz, 1014 MB memory, Windows 7 operating system, and using Java Netbeans 6.9.1 with jdk1.6.0_18.

A. Dataset

This study uses 3 types of datasets. The explanation of the dataset is explained as follows:

Non SPAM training dataset: is a collection of non-spam datase ts totaling 400 data used to get keyterms from non spam files.

SPAM training dataset: is a collection of spam datasets totaling 100 data used to get keyterms from spam files.

Test dataset: is a collection of datasets that are used to test the incoming files. This test data consists of 200 data consisting of 160 Non-SPAM data and 40 SPAM datasets.

B. Testing

Each test conducted will look for accuracy, precision, recall, and F-Measure values using 160 non-spam data and 40 spam data.

1. Testing with Test Data 1

In this test, the accuracy, precision, recall, and F-Measure values will be searched using 100 non-spam email data and 25 spam email data based on keywords from keyword feature extraction from the same SPAM data.

Based on tests carried out using 500 keywords from HAM (Non SPAM) and SPAM from the extraction results using the Fuzzy Association Rule Mining, the accuracy value is 0.984, precision is 0.9253, recall is 1, and F-Measure is 0.961.

2. Testing with Test Data 2

In this test, the accuracy, precision, recall, and F-Measure values will be searched using 200 non-spam email data and 50 spam email data based on keywords from keyword feature extraction from the same SPAM data. Based on tests carried out using 500 keywords from HAM (Non SPAM) and SPAM from the extraction results using the Fuzzy Association Rule Mining, the accuracy value is 0.968, precision is 0.875, recall is 0.98, and F-Measure is 0.924.

3. Testing with Test Data 3

In this test, the accuracy, precision, recall, and F-Measure values will be searched using 300 non-spam email data and 75 spam email data based on keywords from keyword feature extraction from the same SPAM data.

Based on tests carried out using 500 keywords from HAM (Non SPAM) and SPAM from the extraction results using Fuzzy Association Rule Mining, the accuracy value is 0.968, precision is 0.871, recall is 0.987, and F-Measure is 0.925.

4. Testing with Test Data 4

In this test, accuracy, precision, recall, and F-Measure values will be searched using 400 non-spam email data and 100 spam email data based on keywords from keyword feature extraction from the same SPAM data.

Based on tests carried out using 500 keywords from HAM (Non SPAM) and SPAM from the extraction results using the Fuzzy Association Rule Mining, the accuracy value is 0.972, precision is 0.884, recall is 0.99, and F-Measure is 0.934.

Tests on different test data can be shown in table 5.1 which is a table of the effect of the amount of test data on the accuracy, precision, recall, and F-Measure values.

Table I. Table of the effect of the amount of test data on accuracy, precision, recall, and F-Measure values

Data test	Data test1	Data test2	Data test3	Data Test4
Accuracy	0.988	0.967	0.969	0.974
Precision	0.928	0.879	0.876	0.889
Recall	1.000	0.988	0.983	0.996
F-Measure	0.961	0.929	0.920	0.936

Based on table 5.1 it can be illustrated in a graph the effect of the amount of test data on the accuracy, precision, recall, and F-Measure values shown in Figure 2.

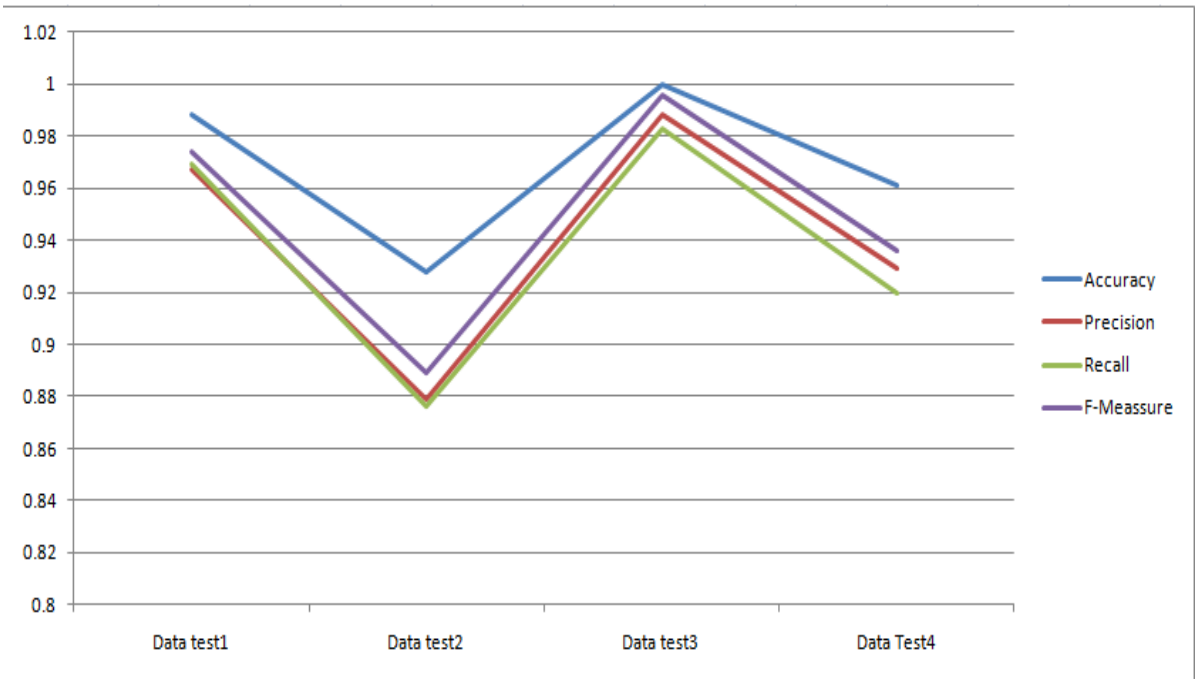

Figure 2. The effect of the amount of test data on the accuracy, precision, recall, and F-Measure values

Based on the tests carried out, it can be seen in Figure 5.1 that the use of test data 1 has the highest value among the others. The high test value of test data 1 occurs because test data 1 has a small amount of data so that the noise that occurs in e-mail detection errors is getting smaller.

IV. CONCLUSION

Based on the tests conducted, the use of the Fuzzy Association Rule Mining method in detecting SPAM and non-SPAM on e-mail is good, this can be seen in the test where the accuracy, precision, recall, and F-Measure values are quite high, which is close to the value of 1.

ISSN 2443-2210 (Print)
ISSN 2443-2229 (Online)

REFERENCES

[1] Appavu, S., et al, 2007. *Association Rule Mining for Suspicious Email Detect ion: A Data Mining Approach*," IEEE.

[2] Rozi, F.,et al, 2015. *Ekstraksi Kata Kunci Berdasarkan Hipernim Menggunakan Fuzzy Association Rule Mining untuk Pengelompokan Dokumen*, J. Ilm. Teknol. Inf., vol. 13, no. 2.

[3] T. T. A. Putri, et al, 2019. *Analysis and Detection of Hoax Contents in Indonesian News Based on Machine Learning*, JIPN (Journal Informatics Pelita Nusantara), vol. 4, no. 1.

[4] B. Santoso, 2019. *An Analysis of Spam Email Detection Performance Assessment Using Machine Learning*, JOIN (Jurnal Online Informatika) Vol.4 no.1.

[5] C. Chen, F. S. C. Tseng, and T. Liang, 2010. *Mining fuzzy frequent itemsets for hierarchical document clustering*, Inf. Process. Manag., vol. 46, no. 2.

[6] N. Alias, et al, 2019, *Video spam comment features selection using machine learning techniques*, Indones. J. Electr. Eng. Comput. Sci., vol. 15, no. 2, pp. 1046–1053.

[7] Dahliar Ananda, 2011. *Pembangunan Aplikasi Pemfilteran Email Spam Dengan Menggunakan Metode Pembeda Markov*, Jurnal Teknologi Informasi Politeknik Telkom Vol. 1, No. 1

[8] A. Prasetyo, et al, 2010. *Klasifikasi Spam Email Dengan Metode Naive Bayes Classifier*, PENS Surabaya.

[9] Rozi, F., and Sukmana, 2016. *Penggunaan moving average dengan metode hybrid artificial neural network dan fuzzy inference system untuk prediksi cuaca*, J. Ilm. Penelit. dan Pembelajaran Inform., vol. 1, no. 2, pp. 38–42

[10] S. Hershkop and S.J. Stolfo, 2009. *Identifying spam without peeking at the contents*, Crossroads, ACM Press, 11(2).

[11] Rozi, F. and Kartadie, R. 2017. *Clustering Dokumen dengan Semantic Word Holonim dan Fuzzy Association Rule Mining, Semnasteknomedia Online*, vol. 5, no. 1.