# An Agent Model using Naïve Bayesian for Email Classification

### Muhammad Hasbi[*)]
Computer Science Post Graduate Program, Faculty of Mathematics and Natural Sciences. Gadjahmada University, Yogyakarta, Indonesia

### Retantyo Wardoyo
Computer Science Post Graduate Program, Program, Faculty of Mathematics and Natural Sciences. Gadjahmada University, Yogyakarta, Indonesia

### Jazi Eko Istiyanto
Computer Science Post Graduate Program, Program, Faculty of Mathematics and Natural Sciences. Gadjahmada University, Yogyakarta, Indonesia

### Khabib Mustofa
Computer Science Post Graduate Program, Program, Faculty of Mathematics and Natural Sciences. Gadjahmada University, Yogyakarta, Indonesia

## ABSTRACT
it is important to carry out email classification to determine its topic [2],[3],[4],[5],[6],[7],[8]. This paper is aimed at making new agents model to determine the email topic by classifying them based on the subject and content autonomously. This domain model is university archiving. The email topic is the keyword of the job description in the university's units. The email target, except the one to the university director, is based on the email topic. The classification method used was Naive Bayesian and Gaussian Density Methods. The agents used were those with proactive characteristic that can work autonomously in classifying emails. The development of this new model results in the detailed email target. Using this model, most emails can be classified correctly according to the categories.

## Keywords
classification, agent, email, naïve bayesian, proactive.

## 1. INTRODUCTION
Email classification has been done such as by filtering emails [1], J48 [2], Support Vector Machine, Boost, Naive Bayesian [3], [4], Random Forest [5], [6]. Email classification using reversal propagation technique [7], and selection method [8] have ever been done as well.

Email classification to filter spam using onthology has ever been done too [2] while the one that is used to help knowing the email content has not been searched. The using of agent based on email can be used as a system to process data [9] has ever been carried out too. On the other hand, the using of agent in classifying and knowing the email's content have not been carried out.

Email that can be called digital Document has also been researched. Document classification using *Hidden Markov Models* [10], web document classification with onthology [11], document classification using agent with statistics methods [12], document's title classification using agent [13] and document classification with word based compression method joined with cluster analysis [14] have been done as well. The comparison between document classification with single agent (machine) and multi-agent in computer network [15] has also been carried out. The result was that the classification using multi-agent is more efficient for the more number of vocabulary.

The goal of this paper is to make a new agent model to classify emails based on their subject and content autonomously. The other goal is to determine the email target. The email domain is on the university's archiving. Email classification was done in this research to get email's topic based on the subject and content of the email. The email topic is the keyword of the job description in the univeristy's Units. The unit of the email target, except to the director, is also determined based on the email topic. The classification methods used in this research were Naive Bayesian and Gaussian Density. The agents used were the proactive ones that could work autonomously in classifying emails. The development of this model is aimed at being implemented in the university's archiving.

So far, there has not been any researches that use agents in archiving system. The development of proactive agent model to process email classification is to determine the email target of certain unit. This is a new proposal in case of agent model.

## 2. PROPOSED METHOD
### 2.1 Model Development
Agent environment will experience changing in the form of incoming email. This will then be responded by a proactive Secretary agent by classifying it [16]. The classification is aimed at determining the email topic. After the email topic has been figured out, the agent will make action to the environment. The action is in the form of notice to the target agent, that is the Chief agent which will determine the email target based on its email topic.

The architecture of the model development is shown in Figure 1.

Figure 1 shows the general description of the proactive agent model development process to carry out the email classification. The incoming email is classified based on its subject and content by a (proactive) secretary agent to determine the email topic. After that, the secretary agent send the classifcation result to the Chief agent.
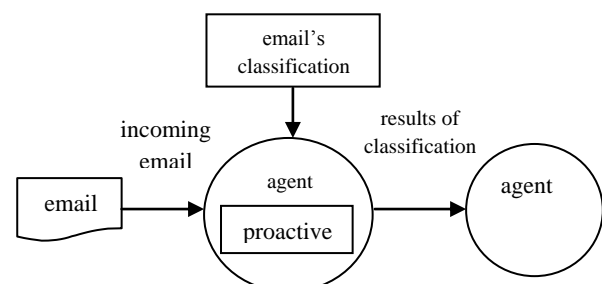


**Fig 1: Agent model in the email classification**

## 2.2 The Development of Agent

The agent's characteristic which was developed was the proactive one in handling the email classification process in the university's email archiving system. The agents involved in this email classification were secretary agent and Chief agent.

The process of the secretary agent's proactive character is as follows: Secretary Agent autonomously uses a sensor to respond the environment changing, that is the presence of the email. The email then is classified by using Naive Bayesian method based on its subject and content. This classification is to determine the topic of the email target. The email topic is a job describtion of each unit at a university. After that, Secretary Agent will react by communicating with the Chief Agent in the form of giving information about the result of the classification and the presence of the email.
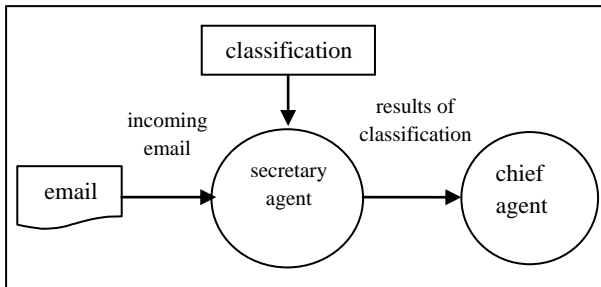


**Fig 2: Agent's proactive character on email's classification**

The email's attribute consists of email's origin, subject, date and time (hour), and the email's content as well. Take a look at Figure 2 for a more specific description. The Secretary Agent respond the incoming email by classifying it which is not only based on its subject but also its content. This classification is carried out to determined the email's topic. The Secretary Agent notifies the Chief Agent to give a follow-up action.

The object used in this research is a university in which its email classification is in the process of email archiving whose its agent being used has its own role. For instance, the Secretary Agent uses its proactive character. The incoming email is responded by the Secretary Agent proactively carries out the process of the email classification. The roles of each agent in the process of the email archiving is seen in Table 1 below.

The email classification is done by the agent. Therefore, the process is autonomous in which the method used in the email classification is Naive Bayesian. The domain of the classification is the email archiving at a university. The classification is done by determining the keywords (feature) of the terms (words) in the university's domain. After that, email topic determination is run. The email topic is gained from the job description of each unit at the university. This email topic is the one that will be the class name.

**Table 1. Agent's role**

| Agent | Character | Structure | Percept | Environment | Actuator | Sensor | Performance measure |
|---|---|---|---|---|---|---|---|
| Secretary Agent | Proactive | Goal-based agent | Coming email | Chief Agent | Sending email's classification | Function of receiving the coming email | The result of email classification based on its content and subject |

The keyword (feature) is chosen from the numerous terms of job description at at the university. Therefore, there must be a limitation for the scope of the keyword domain such as by limiting only on the academic department. In here, a simulation of determining the keyword of several terms is carried out.

The determination of the email topic that will be the class name is also limited to suit the domain of the keyword. Take an example the keyword is the academic department in a university, therefore the class topic is the description of the units' job in it. The number of the class topic is also limited due to the numerous class topic in the academic domain. In short, the class topics to be discussed are just a few.

## 3. EMAIL CLASSIFICATION

Email Classification Method using Naive Bayesian

The stages of email classification:

### 3.1.1 Email data Training [17], [18]

$$p(C_j|K_1,..., K_n) = p(C_j)\, p(K_1,...,K_n|C_j)\, /\, p(K_1,...,K_n). \qquad (1)$$

$$p(K_1, K_2, .. K_n |C_j) = p(K_1|C_j) * p(K_2|C_j)* ...* p(K_n|C_j). \qquad (2)$$

$$p(C_j |K_1, K_2, .. K_n) = p(C_j) * p(K_1, K_2, .. K_n |C_j). \qquad (3)$$

$$p(C_j |K_1, K_2, .. K_n) =$$
$$p(C_j) * (p(K_1|C_j) * p(K_2|C_j)* ...* p(K_n|C_j)\,). \qquad (4)$$

$$C_{nb} = \arg\max c_j \epsilon C\ P(C_j)\ \Pi^n_{i=1}\ P(K_i|C_j). \qquad (5)$$

$K_1, K_2, .. K_n$ = Attribute with the keyword $K_i$.

$p(C_j |K_1, K_2, .. K_n) =$ Probability of C Class$_j$ with the feature (word) $K_1, K_2, .. K_n$.

$p(C_j)$ = Probability of each C class$_j$.

$p(K_1|C_j)$ = Probability of word $K_1$ in $C_j$ Class.

$p(K_2|C_j)$ = Probability of word $K_2$ in $C_j$ Class.

$p(K_n|C_j)$ = Probability of word $K_n$ in $C_j$ Class.

$C_{nb}$ = Classification of Naïve Bayesian by taking the biggest value.

### 3.1.2 Testing of data email

Gaussian Density Function is presented in the following equation [19] :

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

(6)

$f$(x) = gaussian density function.

x  = The number of the word $K_i$.

μ  = The average of the word p($K_1$, $K_2$, .. $K_n$ |$C_j$).

σ  = The deviation standard of p($K_1$, $K_2$, .. $K_n$ |$C_j$).

Based on (4),(5) and (6) we have: (7)

$$Cnb = \arg\max c_j \epsilon C\ P(C_j) \prod_{i=1}^{n}\ [\frac{1}{\sqrt{2\pi\sigma_i}}e^{\frac{-(x_i-\mu_i)^2}{2\sigma_i^2}}] \qquad (7)$$

# 4. RESULT AND DISCUSSION

## 4.1 Agent Model

The agent model is implemented by using Java Netbeans 6.9.1. The constructed agent model is the secretary agent which is proactive in receiving email. It classifies email based on its subject and content. Figure 3 a show description of Secretary Agent.
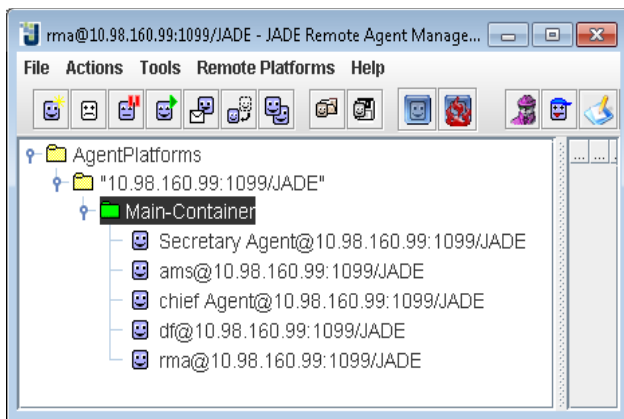


**Fig 3: Secretary agent in email classification**

Figure 3 shows the active agents, Secretary Agent and Chief Agent. The Secretary agent checks emails in the secretary's email server and classifies them. While the Chief Agent receives the result of the email classification from the Secretary Agent and determines the email target.

## 4.2 Two Stages of Email classification

### 4.2.1 Email content data training

Stage 1: form the data training manually by using Naive Bayesian method. In this case, the first thing to do is determining how many email classes that will be formed. For example, there will be 6 email classes going to be formed: Reseach and Service Email, Academic Email, Quality Assurance Email, Student affairs Email, Chief Email and other email.

Stage 2: determining the feature (Key Word) which is used to classify the email's content that is all "words" being used.

For example, the word "student", "research", "student affairs", "accreditation", "foundation", "academic", "study", "lam", "lppm", "research", "position", "assurance", "functional", "service", "grant", "curriculum", "bidikmisi", "institution", "borang", "ban", "tracer", "base", "scholarship", "serdos", "jafa", "kkni", "reviewer", "approach", "instrument", "scp", "extention", "demand".

Next is calculating the statistic parameter such as the μ (mean) dan σ (deviation standard) of each feature in which there are 32 features. The number of emails used is 489 as the data

learning (training). Data learning μ (mean) dan σ (deviation standard) are shown on Table 2 and 3.

**Table 2. Mean**

| Feature (word) | (1) (μ) | (2) (μ) | (3) (μ) | (4) (μ) | (5) (μ) | (6) (μ) |
|---|---|---|---|---|---|---|
| "student" | 0.033 | 0.126 | 0.143 | 2.577 | 0.200 | 0.068 |
| "research" | 2.133 | 0.012 | 0.143 | 0.077 | 0.200 | 0.019 |
| "student affairs" | 0.067 | 0.069 | 0.143 | 1.500 | 0.100 | 0.003 |
| "accreditation" | 0.067 | 0.081 | 1.286 | 0.077 | 0.200 | 0.012 |
| "foundation" | 0.067 | 0.092 | 0.143 | 0.077 | 0.600 | 0.012 |
| "academic" | 0.067 | 0.149 | 0.001 | 0.077 | 0.200 | 0.006 |
| "study" | 0.067 | 0.081 | 0.143 | 0.077 | 0.200 | 0.053 |
| "lam" | 0.033 | 0.126 | 0.071 | 0.039 | 0.100 | 0.012 |
| "lppm" | 0.267 | 0.023 | 0.143 | 0.077 | 0.100 | 0.006 |
| "researcher" | 0.633 | 0.023 | 0.143 | 0.077 | 0.200 | 0.009 |
| "position" | 0.067 | 0.103 | 0.143 | 0.077 | 0.100 | 0.003 |
| "assurance" | 0.067 | 0.035 | 0.500 | 0.077 | 0.100 | 0.003 |
| "functional" | 0.067 | 0.103 | 0.143 | 0.039 | 0.200 | 0.006 |
| "service" | 0.267 | 0.023 | 0.214 | 0.077 | 0.100 | 0.031 |
| "grant" | 0.267 | 0.035 | 0.143 | 0.039 | 0.100 | 0.009 |
| "curiculumn" | 0.067 | 0.092 | 0.143 | 0.077 | 0.100 | 0.006 |
| "bidikmisi" | 0.067 | 0.023 | 0.143 | 0.731 | 0.200 | 0.009 |
| "institution" | 0.067 | 0.081 | 0.643 | 0.039 | 0.200 | 0.025 |
| "borang" | 0.067 | 0.023 | 0.714 | 0.039 | 0.200 | 0.016 |
| "ban" | 0.067 | 0.046 | 0.286 | 0.039 | 0.200 | 0.012 |
| "tracer" | 0.067 | 0.023 | 0.143 | 0.269 | 0.200 | 0.009 |
| "base" | 0.067 | 0.081 | 0.143 | 0.039 | 0.200 | 0.009 |
| "scholarship" | 0.067 | 0.023 | 0.143 | 0.231 | 0.200 | 0.009 |
| "serdos" | 0.067 | 0.058 | 0.143 | 0.077 | 0.200 | 0.009 |
| "jafa" | 0.067 | 0.058 | 0.143 | 0.039 | 0.200 | 0.006 |
| "kkni" | 0.067 | 0.081 | 0.143 | 0.077 | 0.200 | 0.012 |
| "reviewer" | 0.167 | 0.023 | 0.143 | 0.077 | 0.200 | 0.009 |
| "approach" | 0.067 | 0.046 | 0.143 | 0.077 | 0.200 | 0.006 |
| "instrument" | 0.067 | 0.023 | 0.143 | 0.077 | 0.200 | 0.006 |
| "scp" | 0.067 | 0.023 | 0.071 | 0.154 | 0.200 | 0.006 |
| "extention" | 0.067 | 0.058 | 0.143 | 0.077 | 0.200 | 0.034 |
| "demand" | 0.033 | 0.023 | 0.071 | 0.077 | 0.200 | 0.155 |

Table 2 and Table 3 shows the calculation result of the average number of the words and the deviation standard in the email training. The names of the column numbers are column (1)=Class of Research and Service, (2)=Class of Academic, (3)=Class of Assurance, (4) Class of Student Affairs, (5) Class of Senate, (6)=Class of Others.

### 4.2.2 Data Email Testing

For example, there are incoming emails with the data: Emails with the words of "student"=0, "research"=0, "student affairs"=0, "accreditation"=0, "foundation"=0, "academic"=0, "study"=0, "lam"=0, "lppm"=0, "research"=0, "position"=0, "assurance"=0, "functional"=0, "service"=0, "grant"=0, "curriculum"=0, "bidikmisi"=0, "institution"=0, "borang"=0, "ban"=0, "tracer"=0, "base"=0, "scholarship"=0, "serdos"=2, "jafa"=0, "kkni"=0, "reviewer"=0, "approach"=0, and "instrument"=0, "scp"=0, "extention"=0, "demand"=0.

**Table 3. Standard deviation**

| Feature (word) | (1) (σ) | (2) (σ) | (3) (σ) | (4) (σ) | (5) (σ) | (6) (σ) |
|---|---|---|---|---|---|---|
| "student" | 0.183 | 0.367 | 0.363 | 2.580 | 0.422 | 0.328 |
| "research" | 3.491 | 0.107 | 0.535 | 0.272 | 0.422 | 0.176 |
| "student affairs" | 0.254 | 0.452 | 0.363 | 2.267 | 0.316 | 0.056 |
| "acreditation" | 0.254 | 0.488 | 2.585 | 0.272 | 0.422 | 0.136 |
| "foundation" | 0.254 | 0.328 | 0.363 | 0.272 | 1.578 | 0.111 |
| "academic" | 0.254 | 0.540 | 0.001 | 0.272 | 0.633 | 0.079 |
| "study" | 0.254 | 0.463 | 0.535 | 0.392 | 0.422 | 0.354 |
| "lam" | 0.183 | 1.179 | 0.267 | 0.196 | 0.316 | 0.223 |
| "lppm" | 0.740 | 0.151 | 0.535 | 0.272 | 0.316 | 0.112 |
| "researcher" | 1.542 | 0.151 | 0.363 | 0.272 | 0.422 | 0.125 |
| "position" | 0.254 | 0.591 | 0.363 | 0.272 | 0.316 | 0.056 |
| "assurance" | 0.254 | 0.239 | 1.345 | 0.272 | 0.316 | 0.056 |
| "functional" | 0.254 | 0.591 | 0.363 | 0.196 | 0.422 | 0.079 |
| "service" | 0.583 | 0.151 | 0.579 | 0.272 | 0.316 | 0.352 |

| "grant" | 0.692 | 0.184 | 0.363 | 0.196 | 0.316 | 0.096 |
|---|---|---|---|---|---|---|
| "curiculumn" | 0.254 | 0.542 | 0.363 | 0.272 | 0.316 | 0.079 |
| "bidikmisi" | 0.254 | 0.151 | 0.363 | 2.183 | 0.422 | 0.096 |
| "institution" | 0.254 | 0.651 | 1.499 | 0.196 | 0.422 | 0.175 |
| "borang" | 0.254 | 0.151 | 1.069 | 0.196 | 0.422 | 0.185 |
| "ban" | 0.254 | 0.260 | 0.825 | 0.196 | 0.422 | 0.111 |
| "tracer" | 0.254 | 0.151 | 0.363 | 1.373 | 0.422 | 0.096 |
| "base" | 0.54 | 0.463 | 0.363 | 0.196 | 0.422 | 0.096 |
| "scholarship" | 0.254 | 0.151 | 0.363 | 0.710 | 0.422 | 0.125 |
| "serdos" | 0.254 | 0.385 | 0.363 | 0.272 | 0.422 | 0.096 |
| "jafa" | 0.254 | 0.536 | 0.363 | 0.196 | 0.422 | 0.079 |
| "kkni" | 0.254 | 0.410 | 0.363 | 0.272 | 0.422 | 0.111 |
| "reviewer" | 0.913 | 0.151 | 0.363 | 0.272 | 0.422 | 0.096 |
| "approach" | 0.254 | 0.429 | 0.363 | 0.272 | 0.422 | 0.079 |
| "instrument" | 0.254 | 0.214 | 0.535 | 0.272 | 0.422 | 0.079 |
| "scp" | 0.254 | 0.151 | 0.267 | 0.613 | 0.422 | 0.079 |
| "extension" | 0.254 | 0.279 | 0.363 | 0.272 | 0.422 | 0.228 |
| "demand" | 0.183 | 0.151 | 0.267 | 0.272 | 0.422 | 0.720 |

**Table 4. The result of the gaussian function calculation of each class**

| Feature (word) | Σ word | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|---|
| "student" | 0 | 0.918 | 0.620 | 0.613 | 0.151 | 0.549 | 0.682 |
| "research" | 0 | 0.177 | 1.211 | 0.527 | 0.735 | 0.549 | 0.947 |
| "student affairs" | 0 | 0.765 | 0.586 | 0.613 | 0.213 | 0.675 | 1.688 |
| "accreditation" | 0 | 0.765 | 0.564 | 0.219 | 0.735 | 0.549 | 1.077 |
| "foundation" | 0 | 0.765 | 0.670 | 0.613 | 0.735 | 0.295 | 1.191 |
| "academic" | 0 | 0.765 | 0.523 | 0.000 | 0.735 | 0.477 | 1.418 |
| "study" | 0 | 0.765 | 0.578 | 0.527 | 0.625 | 0.549 | 0.664 |
| "lam" | 0 | 0.918 | 0.365 | 0.745 | 0.884 | 0.675 | 0.844 |
| "lppm" | 0 | 0.435 | 1.016 | 0.527 | 0.735 | 0.675 | 1.193 |
| "researcher" | 0 | 0.295 | 1.016 | 0.613 | 0.735 | 0.549 | 1.127 |
| "position" | 0 | 0.765 | 0.511 | 0.613 | 0.735 | 0.675 | 1.688 |
| "assurance" | 0 | 0.765 | 0.808 | 0.321 | 0.735 | 0.675 | 1.688 |
| "functional" | 0 | 0.765 | 0.511 | 0.613 | 0.884 | 0.549 | 1.418 |
| "service" | 0 | 0.471 | 1.016 | 0.490 | 0.735 | 0.675 | 0.670 |
| "grant" | 0 | 0.445 | 0.915 | 0.613 | 0.884 | 0.675 | 1.280 |
| "curiculumn" | 0 | 0.765 | 0.534 | 0.613 | 0.735 | 0.675 | 1.418 |
| "bidikmisi" | 0 | 0.765 | 1.016 | 0.613 | 0.255 | 0.549 | 1.280 |
| "institution" | 0 | 0.765 | 0.491 | 0.297 | 0.884 | 0.549 | 0.945 |
| "borang" | 0 | 0.765 | 1.016 | 0.309 | 0.884 | 0.549 | 0.926 |
| "ban" | 0 | 0.765 | 0.770 | 0.414 | 0.884 | 0.549 | 1.191 |
| "tracer" | 0 | 0.765 | 1.016 | 0.613 | 0.334 | 0.549 | 1.280 |
| "base" | 0 | 0.765 | 0.578 | 0.613 | 0.884 | 0.549 | 1.280 |
| "scholarship" | 0 | 0.765 | 1.016 | 0.613 | 0.449 | 0.549 | 1.127 |
| "serdos" | 2 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| "jafa" | 0 | 0.765 | 0.542 | 0.613 | 0.884 | 0.549 | 1.418 |
| "kkni" | 0 | 0.765 | 0.611 | 0.613 | 0.735 | 0.549 | 1.191 |
| "reviewer" | 0 | 0.411 | 1.016 | 0.613 | 0.735 | 0.549 | 1.280 |
| "approach" | 0 | 0.765 | 0.606 | 0.613 | 0.735 | 0.549 | 1.418 |
| "instrument" | 0 | 0.765 | 0.857 | 0.527 | 0.735 | 0.549 | 1.418 |
| "scp" | 0 | 0.765 | 1.016 | 0.745 | 0.494 | 0.549 | 1.418 |
| "transfer" | 0 | 0.765 | 0.739 | 0.613 | 0.735 | 0.549 | 0.827 |
| "demand" | 0 | 0.918 | 1.016 | 0.745 | 0.735 | 0.549 | 0.459 |

Since the data is continous, the probability value of each P (Feature/Class), is approached by using Gaussian Normal Distribution Probability Formula (6).

The Gaussian Density table is shown in Table 4. Then, the calculation to determine the email classification is done as proposed by the formula (7). The complete result of each classification of the emails is shown in Table 5.

**Table 5. Maximum percentage of each class email**

| Email | Real class | System class | Persentage |
|---|---|---|---|
| Email1 | Research and Service | Research and Service | 89.35% |
| Email2 | Research and Service | Research and Service | 100.00% |
| Email3 | Research and Service | Research and Service | 89.35% |
| Email4 | Others | Others | 100.00% |
| Email5 | Academic | Academic | 99.86% |
| Email6 | Academic | Senate * | 47.32% |
| Email7 | Others | Others | 100.00% |
| Email8 | Others | Others | 100.00% |
| Email9 | Others | Others | 100.00% |
| Email10 | Academic | Senate* | 91.92% |
| Email11 | Student Affair | Student Affairs | 100.00% |
| Email12 | Academic | Senate* | 79.73% |
| Email13 | Others | Others | 100.00% |
| Email14 | Student Affairs | Student Affairs | 55.83% |
| Email15 | Quality Assurance | Quality Assurance | 100.00% |
| Email16 | Others | Others | 100.00% |
| Email17 | Senate | Otehrs* | 100.00% |
| Email18 | Others | Others | 100.00% |
| Email19 | Academic | Others* | 100.00% |
| Email20 | Research and Service | Research and Service | 100.00% |
| Email21 | Senate | Others* | 100.00% |
| Email22 | Research and Service | Research and Service | 100.00% |
| Email23 | Research and Service | Research and Service | 100.00% |
| Email24 | Others | Others | 100.00% |
| Email25 | Others | Others | 100.00% |

The next step is determining the email classification as proposed by the formula (7). The complete result of the classification of each email is tabulated in Table 5. However, Table 5 is only as the representative of the whole 103 email data testings.

As tabulated in Table 5, we can see that the maximum value of the result of the classification formula (7) of each Email testing. The maximum email percentage shows the class of the email. For example, Email1 is classified as email of Research and Service due to its maximum value of 89.35%.

Then, the accuracy of the result of the email classification as tabulated on Table 5 is calculated by using the Confusion Matrix Table [20]. See Table 6 (true positive (tp) = 81, false negative (fn) =17, false negative (fp) = 0, true negative (tn) = 5). In Table 5 we can see that the data with * mark shows the real email class from different system. The numbers are 22 out of 103 email testings.

**Table 6. Confusion matrix**

| System Real | True | False |
|---|---|---|
| True | 81 | 17 |
| False | 0 | 5 |

As seen in Table 6, the accuracy of this system model is (tp+tn/(tp+fn+fp+tn) = 83%.

**Table 7. The Accuracy Comparison**

| Method | Accuracy (%) |
|---|---|
| Classifying emails by combining the results from several classifications ( Grey List Method) [3] | 97 |
| Filtering email spams by using multi-stage classification technique [4] | 97.05 |
| Classifying email spams by using 3-tier [6]. | 96.242 |
| Classifying email by using Back Propagation Technique [7]. | 87 |
| Classifying email by using Data Reduction [8]. | 97 |
| Calssifying email based on subject and content by agent using Naïve Bayesian (being proposed). | 83 |

The accuracy gained by using the proposed agent model is 83%. However, this model classifies emails based on their subject and content done by agent while the other methods do not. The accuracy comparison between the proposed method and the others are seen in Table 7.

## 5. CONCLUSIONS

The development of the new agent model is proposed to determine the email topic with the proactive agent autonomously. The proactive agent carries out email classification based on its content and subject by using Naive Bayesian method. The domain of this email classification is the University. The classification is to determine the email topic. It is the job description of each Unit in the university which then can be used to know which Unit it is actually. We did the experiment by using 489 training emails and 103 testing emails with 6 classes and 32 word fetaures. The Agent Model in the classification is resulting 83% of accuracy.

## 6. ACKNOWLEDGMENT

## 7. REFERENCES

[1] Youn, S. and Mcleod, D., 2007. "Efficient Spam Email Filtering using Adaptive Ontology". *International Conference on Information Technology (ITNG'07) IEEE*.

[2] Taghva, K., Borsack, J., Coombs, J., Condit, A., Lumos, S. and Nartker, T., 2003. "Ontology-based Classification of Email". *Proceedings of the International Conference on Information Technology: Computers and Communications (ITCC'03) IEEE*.

[3] Islam, M.R. and Zhou, W., 2007. "An Innovative Analyser for Email Classification Based on Grey List Analysis". *2007 IFIP International Conference on Network and Parallel Computing - Workshops. IEEE*, pp.178–184.

[4] Islam, M.R. and Zhou, W., 2007. "Email Categorization Using Multi-Stage Classification Technique". *Eighth International Conference on Parallel and Distributed Computing, Applications and Technologies. IEEE*, pp.51–58.

[5] Koprinska, I., Poon, J., Clark, J. and Chan, J., 2007. "Learning to Classify e-mail". *Information Sciences 177 (2007) 2167–2187. Elsevier Inc.*, 177(October 2005), pp.2167–2187.

[6] Islam, M.R., Zhou, W. and Chowdhury, U.M., 2008. "Email Categorization Using (2+1)-tier Classification Algorithms". *Proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science. IEEE*, (1), pp.276–281.

[7] Ayodele, T., Zhou, S. and Khusainov, R., 2010. "Email Classification Using Back Propagation Technique". *International Journal of Intelligent Computing Research (IJICR), Volume 1, Issue 1/2, March/June 2010*, 1(1), pp.3–9.

[8] Islam, R. and Xiang, Y., 2008, "Email Classification Using Data Reduction Method". *Communications and Networking in China (chinacom), 2010 5th International ICST Conference on. IEEE*.

[9] Kwak, J.D., Elmasri, R., Lee, K.K. and Mawr, B., 2000. "Senddata : An Agent for Data *Processing Systems Using Email". Journal Article IEEE, pp.264–270.*

[10] McElroy, J., 2012. "Automatic Document Classification In Small Environments". Faculty of California Polytechnic State University San Luis Obispo, (January).

[11] Fang, J., Guo, L., Wang, X. and Yang, N., 2007. "Ontology-Based Automatic Classification and Ranking for Web Documents Ontology-Based". *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*.

[12] Fu, J., Chang, J,. Huang, Y., Lee, S. and Teng, W., 2007. "Toward an Intelligent Agent System using Document Classification Techniques". *Journal of Computers Vol.18, No.1, April 2007,* pp 31-34.

[13] Song, D,. Bruza, P., Huang, Z. and Lau, R., 2003. "OpenAIR @ RGU The Open Access Institutional Repository at The Robert Gordon University Classifying Document Titles Based on Information Inference". *14th International Symposium, ISMIS 2003 Maebashi City, Japan, October 28-31, 2003 : Proceedings. (ISBN 9783540202561).*

[14] Martinovic, J. and Dvorsky, J., 2007. "Document classification based on The Topic Evaluation and its usage in Data Compression". *2007 Proceedings of IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops*, pp.204–207.

[15] Wooldridge, M., 2009. *An Introduction to MultiAgent Systems* first., Glasgow: John Wiley & Sons.

[16] Peng, S., Mukhopadhyay, S., Raje, R., Palakal, M. and Mostafa, J., "A Comparison Between Single-agent and Multi-agent Classification of Documents". *Proceeding of the 15th International Parallel and Distributed Processing Symposium (IPDPS'01). IEEE*.

[17] Islam M. S., Khaled S. M., Farhan K., Rahman M. A. and Rahman J., 2009, Modeling Spammer Behavior: Naïve Bayes vs. Artificial Neural Networks. Institute of Information Technology, University of Dhaka, Dhaka-1000, Bangladesh. International Conference on Information and Multimedia Technology. IEEE.

[18] pearl, J., 1988, Probabilistic Reasoning In Intelligent Systems: Networks of Plausible Inference. Revised Second Printing. Department of Computer Science University of California Los Angeles. Morgan Kaufmann Publishers, San Mateo, California.

[19] John, G. H., Langley, P., 1995, Estimating Continuous Distributions in Bayesian Classifiers. In proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers. San Matco

[20] Gorunescu F., 2011, Data Mining Concepts, Models and Techniques. Springer-Verlag Berlin Heidelberg.