IOP Conf. Series: Materials Science and Engineering **226** (2017) 012091 doi:10.1088/1757-899X/226/1/012091

Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets

Nurul Fitriah Rusland, Norfaradilla Wahid, Shahreen Kasim, Hanayanti Hafit

Department of Web Technology,
Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia

E-mail: fizzfitty@gmail.com, faradila@uthm.edu.my, shahreen@uthm.edu.my, hanayanti@uthm.edu.my

Abstract.

E-mail spam continues to become a problem on the Internet. Spammed e-mail may contain many copies of the same message, commercial advertisement or other irrelevant posts like pornographic content. In previous research, different filtering techniques are used to detect these e-mails such as using Random Forest, Naïve Bayesian, Support Vector Machine (SVM) and Neutral Network. In this research, we test Naïve Bayes algorithm for e-mail spam filtering on two datasets and test its performance, i.e., Spam Data and SPAMBASE datasets [8]. The performance of the datasets is evaluated based on their accuracy, recall, precision and F-measure. Our research use WEKA tool for the evaluation of Naïve Bayes algorithm for e-mail spam filtering on both datasets. The result shows that the type of email and the number of instances of the dataset has an influence towards the performance of Naïve Bayes.

1. Introduction

Nowadays, e-mail provides many ways to send millions of advertisement at no cost to sender. As a result, many unsolicited bulk e-mail, also known as spam e-mail spread widely and become serious threat to not only the Internet but also to society. For example, when user received large amount of e-mail spam, the chance of the user forgot to read a non-spam message increase. As a result, many e-mail readers have to spend their time removing unwanted messages. E-mail spam also may cost money to users with dial-up connections, waste bandwidth, and may expose minors to unsuitable content. Over the past many years, many approaches have been provided to block e-mail spam [1].

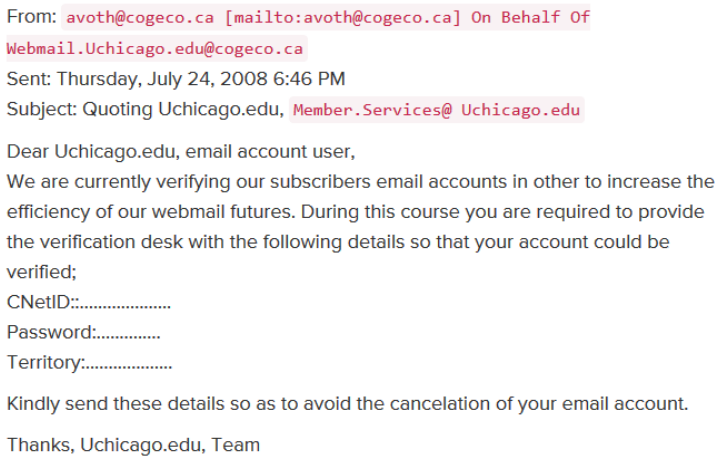
For filtering, some email spam are not being labelled as spam because the e-mail filtering does not detect that email as spam. Some existing problems are regarding accuracy for email spam filtering that might introduce some error. Several machine learning algorithms have been used in spam e-mail filtering, but Naïve Bayes algorithm is particularly popular in commercial and open-source spam filters [2]. This is because of its simplicity, which make them easy to implement and just need short training time or fast evaluation to filter email spam. The filter requires training that can be provided by a previous set of spam and non-spam messages. It keeps track of each word that occurs only in spam, in non-spam messages, and in both. Naïve Bayes can be used in different datasets where each of them has different features and attribute.

The research objectives are: (i) to implement the Naïve Bayes algorithm for e-mail spam filtering on two datasets, (ii) to evaluate the performance of Naïve Bayes algorithm for e-mail spam filtering on the chosen datasets.

The rest of the paper is organized as follows: Section II describes the related work on Naïve Bayes algorithm for e-mail spam filtering. Section III presents the methodology process of e-mail spam using WEKA. Section IV presents the experimental setup. Section V shows the result and analysis on two datasets. Finally, Section VI concludes the work and highlights the direction for future research.

2. Related Work

Spammers are now able to launch large scale spam campaigns, malware and botnets helped spammers to spread spam widely. Upon receiving and opening a spam email, Internet users is exposed to security issues as spams are normally broadcast for bad intention. One of the common email spam example received by users are an email requesting for IDs and passwords(Refer to Figure 1).


From: avoth@cogeco.ca [mailto:avoth@cogeco.ca] On Behalf Of Webmail.Uchicago.edu@cogeco.ca
Sent: Thursday, July 24, 2008 6:46 PM
Subject: Quoting Uchicago.edu, Member.Services@ Uchicago.edu

Dear Uchicago.edu, email account user,
We are currently verifying our subscribers email accounts in other to increase the efficiency of our webmail futures. During this course you are required to provide the verification desk with the following details so that your account could be verified;
CNetID::.....
Password:.....
Territory:.....

Kindly send these details so as to avoid the cancelation of your email account.

Thanks, Uchicago.edu, Team

Figure 1. Sample of spam data requesting for ID and password

Several machine learning algorithms have been employed in anti-spam e-mail spam filtering, including algorithms that are considered top-performers in Text Classification [3], like Boosting algorithm, Support Vector Machines (SVM) algorithm [5] and Naïve Bayes algorithm [7].

Konstantin Tretyakov et al., [6] have evaluated several most popular machine learning methods i.e., Bayesian classification, k-NN, ANNs, SVMs and of their applicability to the problem of spam-filtering. In this work, the author proposed most trivial sample implementation of the named techniques and the comparison of their performance on the PU1 spam corpus dataset is presented. The author used extracting feature to convert all messages to vectors of numbers (feature vectors) and then classify these vectors. This is because most of the machine learning algorithms can only classify numerical objects like vector.

Then the author created the straightforward C++ implementations of the algorithms, and tested them on the PU1 spam corpus. The PU1 corpus consists of 1099 messages, of which 481 are spam. The test setup use by efficiency measure which are precision, legitimate mail fallout and spam fallout. From the result, the performance of the k-nearest neighbours classifier appeared to be poor and the number of false positives was always rather large. According to the author, only the Naïve Bayesian classifier has passed the test.

3. Methodology

This section describes the methodology that is used for the research. The methodology that is used for the filtering method is machine learning techniques that divide by three phases.The methodology is used for the process of e-mail spam filtering based on Naïve Bayes algorithm.

3.1. Naïve Bayes classifier

The Naïve Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combination of values in a given dataset [4]. In this research, Naïve Bayes classifier use bag of words features to identify spam e-mail and a text is representing as the bag of its word. The bag of words is always used in methods of document classification, where the frequency of occurrence of each word is used as a feature for training classifier. This bag of words features are included in the chosen datasets.

Naïve Bayes technique used Bayes theorem to determine that probabilities spam e-mail. Some words have particular probabilities of occurring in spam e-mail or non-spam e-mail. Example, suppose that we know exactly, that the word Free could never occur in a non-spam e-mail. Then, when we saw a message containing this word, we could tell for sure that were spam e-mail. Bayesian spam filters have learned a very high spam probability for the words such as Free and Viagra, but a very low spam probability for words seen in non-spam e-mail, such as the names of friend and family member. So, to calculate the probability that e-mail is spam or non-spam Naïve Bayes technique used Bayes theorem as shown in formula below.

$$P(spam \mid word) = \frac{\mathrm{P(spam).\,P(word|spam)}}{\mathrm{P(spam).\,P(word|spam) + P(non-spam).\,P(word|non-spam)}}$$

Where:

(i) $P(spamword)$ is probability that an e-mail has particular word given the e-mail is spam.
(ii) $P(spam)$ is probability that any given message is spam.
(iii) $P(wordspam)$ is probability that the particular word appears in spam message.
(iv) $P(non-spam)$ is the probability that any particular word is not spam.
(v) $P(wordnon-spam)$ is the probability that the particular word appears in non-spam message.

To achieve the objective, the research and procedure is conducted in three phases. The phases involved are as follows:

(i) Phase 1: Pre-processing
(ii) Phase 2: Feature Selection
(iii) Phase 3: Naïve Bayes Classifier

The following sections will explain the activities that involve in each phases in order to develop this project. Figure 2 shows the process for e-mail spam filtering based on Naïve Bayes algorithm.

3.2. Pre-processing

Today, most of the data in the real world are incomplete containing aggregate, noisy and missing values [9]. Pre-processing of e-mails in next step of training filter, some words like conjunction words, articles are removed from email body because those words are not useful in classification.

As mentioned earlier, we are using WEKA tool to facilitate the experiments. For both experiments, the datasets are presented in Attribute-Relation File Format(ARFF) file (Refer to Figure 3 for sample of data).

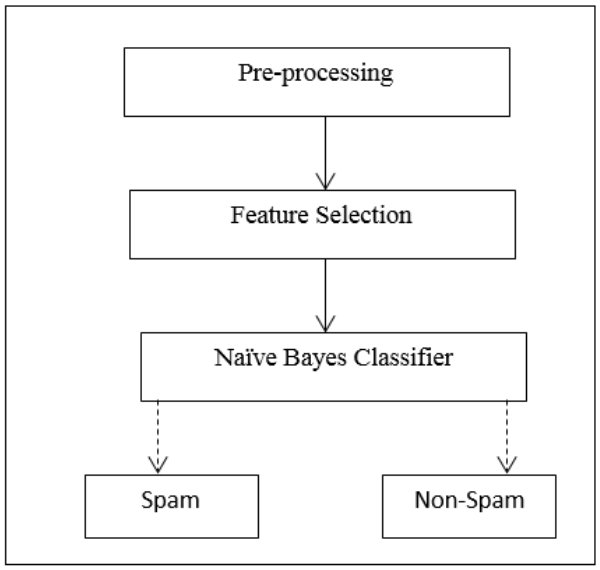


Figure 2. Process of E-mail spam filtering based on Naïve Bayes Algorithm

```
@attribute word_freq_conference   numeric
@attribute char_freq_;            numeric
@attribute char_freq_(            numeric
@attribute char_freq_[            numeric
@attribute char_freq_!            numeric
@attribute char_freq_$            numeric
@attribute char_freq_#            numeric
@attribute capital_run_length_average numeric
@attribute capital_run_length_longest numeric
@attribute capital_run_length_total   numeric
@attribute spam                   {0,1}

@data

0,0.64,0.64,0,0.32,0,0,0,0,0,0,0.64,0,0,0,0.32,0,1.29,1.93,0,0.96,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0.21,0.28,0.5,0,0.14,0.28,0.21,0.07,0,0.94,0.21,0.79,0.65,0.21,0.14,0.14,0.07,0.28,3.47,0,1.59,0,0.43,0
0,0.372,0.18,0.048,5.114,101,1028,1
0.06,0,0.71,0,1.23,0.19,0.19,0.12,0.64,0.25,0.38,0.45,0.12,0,1.75,0.06,0.06,1.03,1.36,0.32,0.51,0,1.16,(
0,0.01,0.143,0,0.276,0.184,0.01,9.821,485,2259,1
0,0,0,0,0.63,0,0.31,0.63,0.31,0.63,0.31,0.31,0.31,0,0,0.31,0,0,3.18,0,0.31,0,0,0,0,0,0,0,0,0,0,0,0,(
0,0,0,0,0.63,0,0.31,0.63,0.31,0.63,0.31,0.31,0.31,0,0,0.31,0,0,3.18,0,0.31,0,0,0,0,0,0,0,0,0,0,0,0,(
0,0,0,0,1.85,0,0,1.85,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,1.92,0,0,0,0,0.64,0.96,1.28,0,0,0,0.96,0,0.32,3.85,0,0.64,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,1.88,0,0,1.88,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0.15,0,0.46,0,0.61,0,0.3,0,0.92,0.76,0.76,0.92,0,0,0,0,0,0.15,1.23,3.53,2,0,0,0.15,0,0,0,0,0,0,0,0.15,
744,445,1257,1
```

Figure 3. Sample of spam data fragment in ARFF format

A full list of the attributes in this data set appears in the "Attributes" frame as shown in Figure 4. Random selection of attribute are performed for the further process.

Attributes *capital_run_length_average*,*capital_run_length_longest* and *capital_run_length_total* are removed from the list by checking the box to their left and hitting the Remove button.

3.3. Feature Selection

After the pre-processing step, we apply the feature selection algorithm, the algorithm which deploy here is Best First Feature Selection algorithm[12].

4. Experimental Setup

The experimental setting of the research is like follows:

4.1. The Evaluation Metric

Evaluation metrics are used to evaluate the performance of WEKA tool based on two datasets that had been chosen. The most simple measure is filtering accuracy namely percentage of

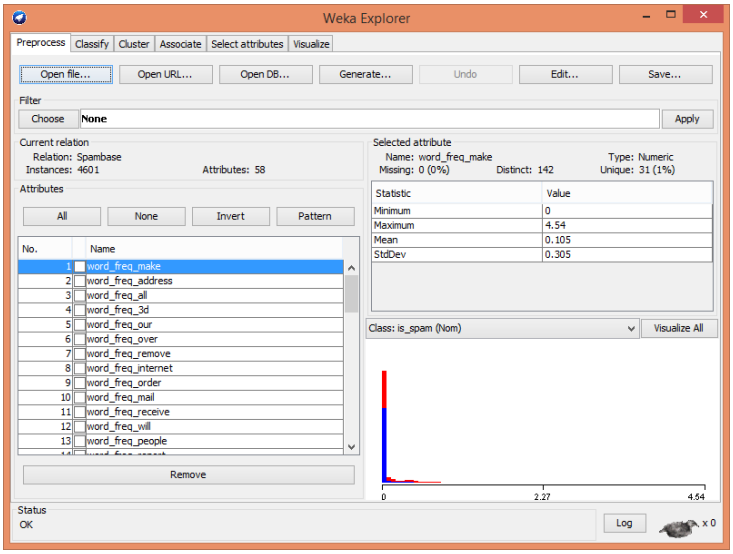


Figure 4. Sample list of the attributes in the "Attributes" frame

messages classified correctly [5]. Table 1 shows the evaluation measures for spam filters.

Table 1. Evaluation measures for spam filters

Evaluation Measure	Evaluation Function
Accuracy	$Acc = \frac{TN+TP}{TP+FN+FP+TN}$
Recall	$r = \frac{TP}{TP+FN}$
Precision	$P = \frac{TP}{TP+FP}$
F-measure	$F = \frac{2pr}{p+r}$

Where accuracy, recall, precision, F-measure, FP, FN, TP and TN are defined as follows:

(i) Accuracy: Percentage of correctly identified spam and not spam message
(ii) Recall: Percentage spam message manage to block
(iii) Precision: Percentage of correct message for spam e-mail
(iv) F-measure: Weighted average of precision and recall
(v) False Positive Rate (FP): The number of misclassified non spam emails
(vi) False Negative Rate (FN): The number of misclassified spam emails
(vii) True Positive (TP): The number of spam messages are correctly classified as spam
(viii) True Negative (TN): The number of non-spam e-mail that is correctly classified as non-spam

4.2. Dataset

Dataset is a collection of data or related information that is composed for separate elements. A collection of dataset for e-mail spam contains spam and non-spam messages. In this research, two datasets are be used to evaluate the performance of Naïve Bayes algorithm to filter e-mail spam.

(i) *Dataset 1: Spam Data*

Spam Data [7] is used in order to test the performance of spam filter based on Naïve Bayes algorithm. This dataset contains 9324 e-mails and 500 attributes. Most of the attributes represent the frequency of a given word or character in the email that corresponds to the instance. This dataset is collected from Usenet posts that exist in the 20 Newsgroup collections and collect from many account e-mails located on different e-mail servers.

(ii) *Dataset 2: SPAMBASE*
SPAMBASE was taken from UCI machine learning repository [8] and was created by Mark Hopkins, Erik Reeber, George Forman, and Jaap Suermondt. This dataset contains 4601 e-mail messages and 58 attributes. This dataset collection of non-spam email came from filled work, personal e-mail and single e-mail account. This dataset is composed of a selection of mail messages, suitable for use in testing spam filtering systems. Each instance in SPAMBASE consists of 58 attributes. Most of the attributes represent the frequency of a given word or character in the email that corresponds to the instance.

5. Result and discussions

This section discussed the experimental result by utilising WEKA tool using Naïve Bayes algorithm. The two datasets are compared based on the percentage of correctly identified spam and non-spam message, percentage of spam message manage to block, percentage of correct message for spam e-mail and weighted average of precision and recall.

For each dataset, 10 run of experiments were conducted using WEKA. The experiments have been done in two parts; (i) using random number of attribute, (ii) using same number of attribute. Figure 5 to 8 are intended for the first part, while Figure 9 is for the second part of the experiment.

For FP, FN, TP and TN, the average for each dataset are as follows:
Spam Data:

- FP: Total 407 number of misclassified non spam emails
- FN: Total 420 number of misclassified spam emails
- TP: Total 1967 number of spam messages are correctly classified as spam
- TN: Total 6530 number of non-spam e-mail that is correctly classified as non-spam

SPAMBASE:

- FP: Total 369 number of misclassified non spam emails
- FN: Total 434 number of misclassified spam emails
- TP: Total 2767 number of spam messages are correctly classified as spam
- TN: Total 1031 number of non-spam e-mail that is correctly classified as non-spam

For the accuracy average, the difference total of two datasets is 8.59% which Spam Data get 91.13% while SPAMBASE get 82.54%(Refer to Figure 5). Figure 6 on the other hand shows that SPAMBASE get the highest percentage with 88% while Spam Data 83% for the average of precision. It means SPAMBASE get almost correctly prediction for spam e-mail. From Figure 7, SPAMBASE get the highest percentage of spam e-mail manage to block than Spam Data with 86%. Lastly from Figure 8, SPAMBASE again get the highest percentage with 87% while Spam Data 83%.

As mention earlier, the second part of the experiment conducted by selecting same number of attributes for each dataset. From Figure 9, we can see Spam Data again achieved the highest percentage for accuracy but still low for precision, recall and F-measure than SPAMBASE. For accuracy, Spam Data get 82.88% than SPAMBASE only get 72.57%. For precision Spam Data get 74% while SPAMBASE get 76%. For recall, the difference total of two datasets is 28% which

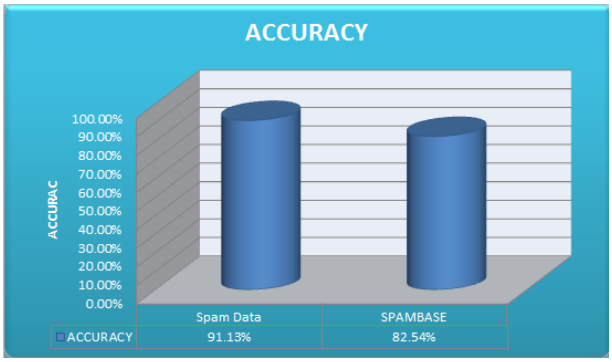


Figure 5. Average Accuracy result for 10 runs of experiment

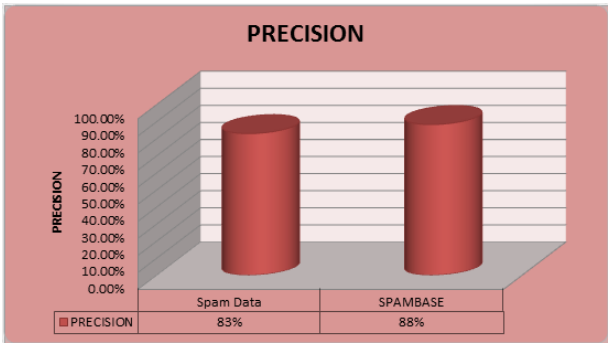


Figure 6. Average Precision result for 10 runs of experiment

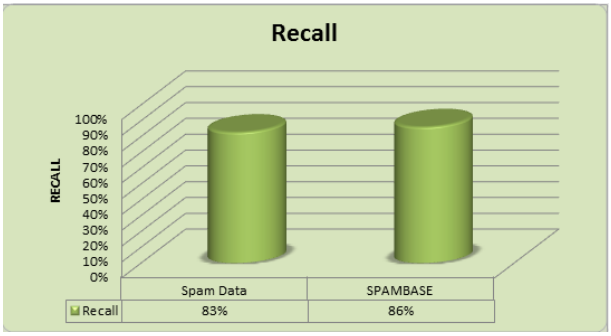


Figure 7. A geRecall result for 10 runs of experiment

Spam Data get only 51% while SPAMBASE get 79%. Lastly, for F-Measure, Spam Data get 60% while SPAMBASE get 77%. As can see, all the results have not much difference from result above. But, from this result, can see the big different total percentage of two datasets than earlier result.

From result that we have been record, we can analyse that Naïve Bayes classifier performs better on SPAMBASE as compared to the Spam Data, even though Spam Data manage to achieve good result on accuracy. This is because filter that gets good result based on precision, recall and F-measure are also important. Spam Data has many attribute and instance of e-mail with overall 9324 e-mails and 500 attributes and Spam Data not manage to performance well with the implement Naïve Bayes classifier. For all the evaluation of metric which precision, recall and F-measure, SPAMBASE performance very well than Spam Data. Even SPAMBASE

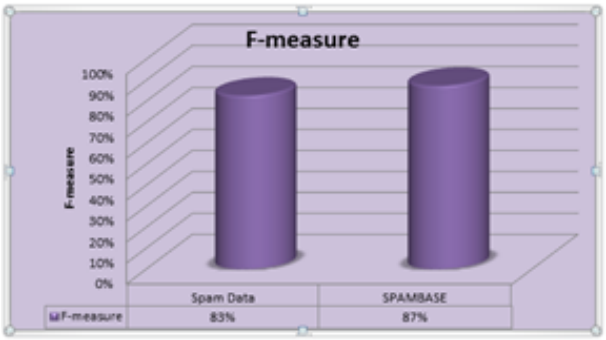


Figure 8. Average F-measure result for 10 runs of experiment

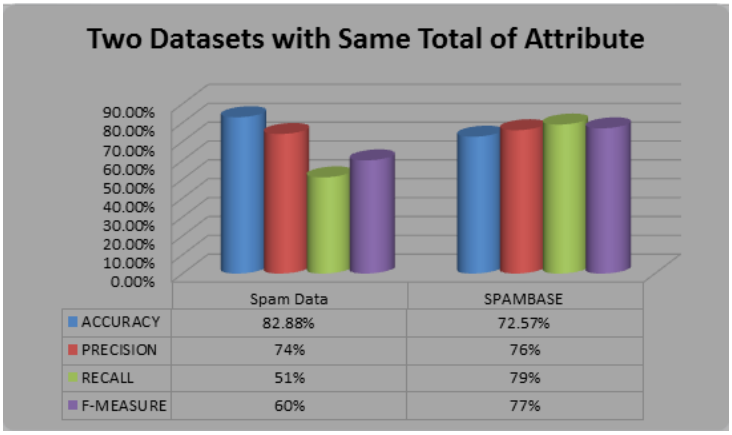


Figure 9. Two Datasets with Same Total Atttribute

dataset only has 4601 instance e-mails and 58 attributes, but Naïve Bayes classifier manage to get good result from this dataset. This is because Naïve Bayes classifier no needs many instances of e-mails and attributes to train the classifier for e-mail spam filtering.

SPAMBASE dataset is multivariate dataset contain data from a single e-mail account while Spam Data dataset collect from many e-mail account. From this we can know, Naïve Bayes classifier can perform well with dataset that come from a single e-mail account than many e-mail account. This is because Naïve Bayes classifier can focus train with various type of e-mail spam that come from single e-mail account located on same e-mail servers.

Besides that, we also test if these datasets have same total of attribute and minimizing the attribute with six attribute, wether it gives different result. From what we get, the result not given different result because Naïve Bayes classifier that used SPAMBASE dataset still has the best performance than Spam Data dataset. But the percentage for accuracy, precision, recall and F-measure drop a little bit. This is proving that total of attribute give important role to Naïve Bayes classifier for filtering e-mail spam.

Although minimizing of attribute decreased the performance of Naïve Bayes classifier, but in the other hand it improved time complexity. The time taken to build model is faster with less attribute and the best performance Naïve Bayes classifier that used SPAMBASE dataset need 0.14 second to build the model. However, it is not important because e-mail spam filtering must have highest accuracy, precision, recall and F-measure to filter that e-mail spam or non-spam.

6. Conclusion

E-mail spam filtering is an important issue in the network security and machine learning techniques; Naïve Bayes classifier that used has a very important role in this process of filtering e-mail spam. The quality of performance Naïve Bayes classifier is also based on datasets that used. As can see, dataset that have fewer instances of e-mails and attributes can give good performance for Naïve Bayes classifier. Naïve Bayes classifier also can get highest precision that give highest percentage spam message manage to block if the dataset collect from single e-mail accounts. So we can see, why performance of Naïve Bayes classifier is good when used SPAMBASE dataset.

Acknowledgement

This research was partially supported by Gates IT Solution Sdn Bhd and Research Grant Vot. U540.

References

[1] Rushdi, S. and Robet, M, "Classification spam emails using text and readability features", *IEEE 13th International Conference on Data Mining*, 2013.
[2] Androutsopoulos, I., Paliouras, G., and Michelakis, "E. Learning to filter unsolicited commercial e-mail", *Technical report NCSR Demokritos*, 2011.
[3] Rathi, M. and Pareek, V. "Spam Mail Detection through Data Mining A Comparative Performance Analysis", *I.J. Modern Education and Computer Science*, 2013, 12, 31-39.
[4] Patil, T. and Sherekar, S. "Performance Analysis of Naïve Bayes and Classification Algorithm for Data Classification", *International Journal Of Computer Science And Applications*, 2013.
[5] Kumar, S., Gao, X., Welch, I. and Mansoori, M., "A Machine Learning Based Web Spam Filtering Approach", IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), Crans-Montana, 2016, pp. 973-980.
[6] Tariq, M., B., Jameel A. Tariq, Q., Jan, R. Nisar, A. S., "Detecting Threat E-mails using Bayesian Approach", *IJSDIA International Journal of Secure Digital Information Age*, Vol. 1. No. 2, December 2009.
[7] Feng, W., Sun, J., Zhang, L., Cao, C. and Yang,Q., "A support vector machine based naive Bayes algorithm for spam filtering," 2016 IEEE 35th International Performance Computing and Communications Conference (IPCCC), Las Vegas, NV, 2016, pp. 1-8.
[8] Tretyakov, K. Machine learning techniques in spam filtering: Data Mining Problem-oriented Seminar, MTAT.03.177, May 2004.
[9] ML & KD- Machine Learning & Knowledge Discovery Group. http://mlkd.csd.auth.gr/concept_drift.html.
[10] UCI Machine Learning Repository Spambase Dataset. University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml/datasets
[11] Kishore, R. K., Poonkuzhali, G. and Sudhakar, P. "Comparative Study on Email Spam Classifier using Data Mining Techniques", *Proceedings of the International MultiConference of Engineers and Computers Science Scientists*, 2012.
[12] Rizky, W. M., Ristu, S., Afrizal, D. "The Effect of Best First and Spreadsubsample on Selection of a Feature Wrapper With Naïve Bayes Classifier for The Classification of the Ratio of Inpatients". Scientific Journal of Informatics, Vol. 3(2), p. 41-50, Nov. 2016.