JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 8. NO. 2 FEBRUARY 2023
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v8i2.2463

COMPARISON OF DIFFERENT KERNEL FUNCTIONS OF SVM CLASSIFICATION METHOD FOR SPAM DETECTION

AAIN Eka Karyawati[1*)]; Komang Dhiyo Yonatha Wijaya[2]; I Wayan Supriana[3]

Informatics Study Program
FMIPA Universitas Udayana
https://if.unud.ac.id
eka.karyawati@unud.ac.id[1]; komangdhiyo66@gmail.com[2]; wayan.supriana@unud.ac.id[3]

(*) Corresponding Author

Abstract—*Today, the use of e-mail, especially for formal online communication, is still often done. There is one common problem faced by e-mail users, which is the frequent receiving of spam messages. Spam messages are generally in the form of advertising or promotional messages in bulk to everyone. Of course this will cause inconvenience for people who receive the SPAM message. SPAM e-mails can be interpreted as junk messages or junk mail. So that spam has the nature of sending electronic messages repeatedly to the owner of the e-mail. This is abuse of the messaging system. One way to solve the spam problem is to identify spam messages for automatic message filtering. Several machine learning based methods are used to classify spam messages. In this study, a comparison was made between several kernel functions (i.e., linear, degree 1 polynomial, degree 2 polynomial, degree 3 polynomial, and RBF) of the SVM method to get the best SVM model in identifying spam messages. The evaluation results based on the Kaggle 1100 dataset showed that the best model were the SVM model with a linear kernel function and a degree 1 polynomial, where both models returned Precision = 0.99, Recall = 0.99, and F1-Score = 0.98. On the other hand, the RBF kernel produced lower performance in terms of Precision, Recall, and F1-Score of 0.95, 0.95, and 0.94, respectively.*

Keywords*: Spam, SVM, Kernel Function, Classification.*

Intisari—*Dewasa ini, penggunaan e-mail, khususnya untuk komunikasi formal secara online, masih sering dilakukan. Ada satu masalah umum yang dihadapi oleh pengguna e-mail, yaitu seringnya menerima pesan spam. Pesan spam umumnya berupa pesan iklan atau promosi secara massal kepada semua orang. Tentu hal ini akan menimbulkan ketidaknyamanan bagi orang yang menerima pesan SPAM tersebut. e-mail SPAM dapat diartikan sebagai pesan sampah atau junk mail. Sehingga spam memiliki sifat mengirimkan pesan elektronik secara berulang-ulang kepada pemilik e-mail tersebut. Ini adalah penyalahgunaan sistem pesan. Salah satu cara untuk mengatasi masalah spam adalah dengan mengidentifikasi pesan spam untuk pemfilteran pesan otomatis. Beberapa metode berbasis pembelajaran mesin digunakan untuk mengklasifikasikan pesan spam. Pada penelitian ini dilakukan perbandingan antara beberapa fungsi kernel (yaitu linear, polinomial derajat 1, polinomial derajat 2, polinomial derajat 3, dan RBF) dari metode SVM untuk mendapatkan model SVM terbaik dalam mengidentifikasi pesan spam. Hasil evaluasi berdasarkan dataset Kaggle 1100 menunjukkan bahwa model terbaik adalah model SVM dengan fungsi kernel linier dan polinomial derajat 1, dimana kedua model mengembalikan Precision = 0.99, Recall = 0.99, dan F1-Score = 0.98. Di sisi lain, kernel RBF menghasilkan kinerja yang lebih rendah dalam hal Precision, Recall, dan F1-Score masing-masing sebesar 0,95, 0,95, dan 0,94.*

Kata Kunci*: Spam, SVM, Fungsi Kernel, Klasifikasi..*

INTRODUCTION

Today, the use of e-mail, especially for formal online communication, is still often done. There is one common problem faced by e-mail users, which is the frequent receiving of spam messages. Spam messages are generally in the form of advertising or promotional messages in bulk to everyone.According to J. Clement as of December 2019 the number of spam e-mails covered 57.26% of the total number of e-mails. Spam is often done for advertising, to get people who are spammed to reply to the message, or to annoy people who are spamming. For this reason, an identification of spam is needed to filter out Spam.

Identification is a specifik task of classification. One approach that usually used to do classification is the machine learning-based method. There have been many studies used machine learning-based spam classification [1]–[6].

A machine learning has the advantage that it is easy to implement and good for high-dimensional data. However, it has the disadvantage of requiring unbiased and large amounts of data. In addition, adjusting parameters and complexity of the model is needed to select the best model.

In this research, a machine-learning based methode namely SVM is used to classify e-mail into two classes (i.e., Spam and not Spam). The aim of this reasearh is to select the best SVM model by comparing some kernel function of the SVM method, besides the parameters.

Several studies [7]–[10] used SVM for spam classification. [7] compared KNN, linear kernel SVM and RBF kernel SVM method. In this study, it was found that the KNN method at k=3 produced the best accuracy of 92.28% while the best accuracy in the SVM method was obtained using the SVM linear kernel with an accuracy of 96.6%. It can be concluded that the SVM method is better than KNN. [8] compared the Naive Bayes method and the SVM method with the RBF kernel to identify Instagram comment spam. The results showed that the SVM method produced an accuracy of 78.49%, which is better than the Naive Bayes method which produced an accuracy of 77.25%.

Other study [9] proposed a combination of KNN and SVM method. It used KNN-based sampling strategy to find close neighbors to improve the performance of the SVM method. The results of the study based on publicly available dataset (Dredze) showed the accuracy increased to about 98%. [10] proposed a new spam detection method that effective in distinguishing spam from its content. During classifying the dataset, the proposed classifier obtained a classification accuracy of 95.32 percent.

In this study several kernel functions (i.e., Linear, Polynomial, and RBF) were investigated to obtain the best SVM model for classifying Spam e-mails. Some experiments were conducted to determine the effect of parameter changes for each kernel function. The SVM performance was measured using the Precision, Recall, and F-Measure metrics.

MATERIALS AND METHODS

Research Steps

The general flow of this research starts from collecting raw data, then data preprocessing (i.e., tokenization, case folding, stop word removal, and stemming), then TF-IDF weighting, training each SVM with various kernels, and finally, evaluating the best SVM. model for each kernel.

Data Collection

The e-mail dataset was collected from https://www.kaggle.com/datasets/venky73/spam-mails-dataset which contains both spam and non-spam e-mails extracted from the e-mail body. The total number of 1100 data was 550 Spam data and 550 non-Spam data, of which 1000 data were used as training data (i.e., to select the best model of each kernel), and 100 data are used for testing (i.e., to evaluate the best model). The feature data used was word frequency. The data was in the form of text in .csv format. An example of a non-Spam e-mail and a Spam e-mail can be seen in Figure 1 and Figure 2, respectively.

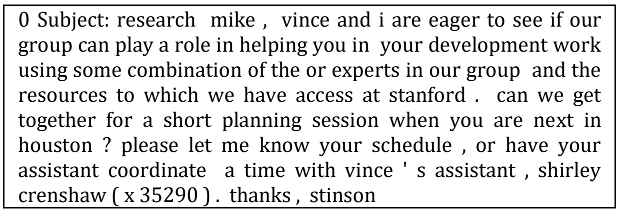
0 Subject: research mike , vince and i are eager to see if our group can play a role in helping you in your development work using some combination of the or experts in our group and the resources to which we have access at stanford . can we get together for a short planning session when you are next in houston ? please let me know your schedule , or have your assistant coordinate a time with vince ' s assistant , shirley crenshaw (x 35290) . thanks , stinson

Figure 1. An Example of Non-Spam e-mail

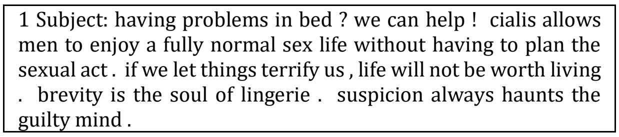
1 Subject: having problems in bed ? we can help ! cialis allows men to enjoy a fully normal sex life without having to plan the sexual act . if we let things terrify us , life will not be worth living . brevity is the soul of lingerie . suspicion always haunts the guilty mind .

Figure 2. An Example of Spam e-mail

TF-IDF Weighting

The term weighting with TF-IDF starts by calculating the term frequency (tf). After that, a temporary weight calculation is carried out with equation (1) and Wf is obtained. To reduce the value of terms that occur frequently, the document frequency (df) of the term is calculated, followed by the calculation of the inverse-document frequency (idf) by equation (2). To get the tf-idf weight, the calculation is carried out according to equation (3) and the result is a term weight matrix. The results of this weighting will be used in the classification process with the SVM method.

Calculate term frequency (tf) by calculating the frequency of terms in the document and term weight (Wf) .

$$Wf_{t,d} = \begin{cases} 1 + logtft,d, jikaf_{t,d} > 0 \\ 0, jikaf_{t,d} \leq 0 \end{cases} \quad \text{.......................} (1)$$

With Wft,d is for weigh term t in document d, tf is for term frequency.

Calculate the document frequency (df) by calculating the frequency of the document where the term is located. Calculate the inverse document frequency (idf) [11].

VOL. 8. NO. 2 FEBRUARY 2023
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v8i2.2463

idf$_t$ = $log \frac{N}{dft}$.. (2)

With idf$_t$ is the inverse document frequency for term t, df$_t$ is document frequency of term t, and N is total frequency of documents.
Calculated the value of TF-IDF weighting [11].

Wtf-idf$_{t,d}$ = $Wf_{t,d} * idf_t$.. (3)

With Wtf-idf is the value of tf-idf weighting, Wf is the weight of term, and idf is inverse document frequency.

Support Vector Machine (SVM)

The Support Vector Machine (SVM) algorithm is a supervised learning method that produces an input-output mapping function from a series of training data that already has a label [12], [13]. Nonlinear kernel functions are often used to convert the input data to a high-dimensional feature space where the input data becomes more separable than the original input space. The algorithm of SVM used in this research is shown in Figure 3.

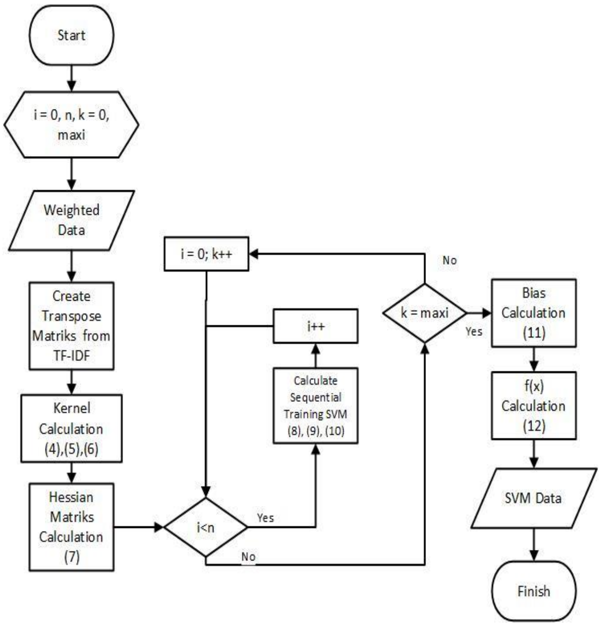
Start
i = 0, n, k = 0, maxi
Weighted Data
Create Transpose Matriks from TF-IDF
Kernel Calculation (4),(5),(6)
Hessian Matriks Calculation (7)
i<n
Yes
No
Calculate Sequential Training SVM (8), (9), (10)
i++
i = 0; k++
k = maxi
No
Yes
Bias Calculation (11)
f(x) Calculation (12)
SVM Data
Finish

Figure 3. The flowchart of SVM

The SVM kernel used in this study were a polynomial, linear and RBF kernel using equation (4), (5), and (6), respectively [14].
Linear Kernel
$K(x_i, x_j) = x_i^T x_j$ .... (4)
Polynomial Kernel

K (xi, xj) = $x_i^T x_j + C^d$ .... *(5)*
RBF Kernel

$K(x_i, x_j) = exp\left(-\gamma |x_i - x_j|^2\right), \gamma > 0$.............. (6)

With K(x$_i$, x$_j$) is kernel fuction, x$_i$ is i-th data, x$_j$ is j-th data, C for slack variable, d for degree, and γ for learning rate.
The steps in using the SVM method are as follows [6]:

a. Initiation of parameters used such as λ and γ (error rate).
b. Calculate the Hessian matrix.

D$_{ij}$ = $y_i y_j \left(K(x_i, x_j) + \lambda^2\right)$.. (7)

With D$_{ij}$ is Hessian matrix value, y$_i$ is i-th class, y$_j$ is j-th class, and λ for error control.

c. Starting from the 1st data to the nth data, do the calculation iterations.

ε$_i$ = $\sum_{j=1}^{n} \alpha_j D_{ij}$ .... (8)
δα$_i$ = $min\{max[\gamma(1 - \varepsilon_i), -\alpha_i], C - \alpha_i\}$............. (9)
α$_i$ = $\alpha_i + \delta\alpha_i$ .... (10)

With ε is error value and α$_i$ is support vector.

$$b = -\frac{1}{2}\left[\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x^-)\right) + \left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x^+)\right)\right]$$

d. From the previous calculation, the largest value of *α$_i$* is sought and calculations are carried out to determine the bias.

.... (11)

With b is for bias value.

e. To find out the results of the sentiment analysis, the f(x) function is calculated.

f(x) = $\sum_{i=0}^{n} \alpha_i y_i K(x_i, x) + b$ .... (12)

With f(x) is for classification function.

Selecting the Best Model

Model selection was done by adjusting the SVM parameters, namely learning rate (γ), error control (λ), and d (polynomial degree), for each kernel. The best model was determined by measuring the model's performance (i.e., F-Measure). The highest F-Measure in each particular combination of kernel parameters was chosen as the best model for that kernel.

Evaluation

The best models of each kernel were evaluated based on 100 testing data. The evaluation metrics used were Precision, Recall, and F-Measure [15].

RESULTS AND DISCUSSION

The SVM Model Selection for Liner Kernel

The Experiments of λ Changes

Table 1 showed the results of the effect of λ changes to the results of the SVM classifier performance with linear kernel. The experiment was carried out with γ = 0.1. The B value was started from 0.01 and progressed to a value of 2. As can be seen, there were no changes in term of Precision, Recall, and F-Measure or the value of B has no effect on the linear kernel.

Table 1. Effect of λ Changes in Linear Kernel

λ	Precision	Recall	F-Measure
0,01	0.9913	0.991	0.9909
0,1	**0.9913**	**0.991**	**0.9909**
1	0.9913	0.991	0.9909
2	0.9913	0.991	0.9909

The Experiments of γ Changes

Table 2 showed the results of testing the effect of changing parameter γ on the results of SVM classifier performance with a linear kernel. The test was carried out with λ = 0.1 and γ started from 0.0001 which was continued until γ = 0.1. As can be seen in Table 2, there was no change in the value of the performance measure (ie, Precision, Recall or N-Measure). So, similar to change B, change γ has no effect on the linear kernel.

Table 2. The Effect of γ Changes in Linear Kernel

γ	Precision	Recall	F-Measure
0,0001	0.9913	0.991	0.9909
0,001	0.9913	0.991	0.9909
0,01	0.9913	0.991	0.9909
0,1	**0.9913**	**0.991**	**0.9909**

The Classification Result of SVM with Polynomial Kernel

The following were the results obtained from the classification using the SVM with polynomial kernel. The experiments was conducted using the 10-fold cross validation method.

The SVM Model Selection for Degree 1 Polynomial Kernel

The Experiments of λ Changes

Table 3 showed the results of testing the effect of λ canges to the results of the SVM classifier performance with a degree 1 polynomial kernel. The testing was carried out with γ = 0.1 and λ was started from 0.01 which continued until λ = 2. From Table 3, it can be seen that there were no change in the value of Precision, Recall and F-measure. The λ values has no effect on polynomial kernel of degree 1.

Table 3. Effect of λ Changes in degree 1 polynomial kernel

λ	Precision	Recall	F-Measure
0,01	0.9923	0.992	0.9919
0,1	**0.9923**	**0.992**	**0.9919**
1	0.9923	0.992	0.9919
2	0.9923	0.992	0.9919

The Experiments of γ Changes

Table 4 showed the results of testing, observing the effect of γ canges to the results of the SVM classifier with a degree 1 polynomial kernel. The test was carried out with λ = 0.1, and γ was started from 0.0001 which continued until 0.1. The best results was found at γ = 0.001 with an average f-measure value of all folds of 0.9919 or 99.19%. As can be seen in Figure 4, the greater the value of γ (learning rate), the lower the value of SVM performance, the evaluation results increase to a peak at 0.001, where after that it decreases. In addition, experiments also showed that a learning rate that was too small gived poor results.

Table 4. Effect of γ Changes in degree 1 polynomial kernel

γ	Precision	Recall	F-Measure
0,0001	0.942	0.933	0.9324
0,001	**0.9931**	**0.993**	**0.992**
0,01	0.9923	0.992	0.9919
0,1	0.9913	0.991	0.9909

The SVM Model Selection for Degree 2 Polynomial Kernel

The Experiments of λ Changes

Table 5 showed the results of the research on the effect of the value of γ changes to SVM classifier performance with a degree 2 polynomial kernel. The experiments was conducted with γ = 0.1 and λ was started from 0.01 which continued to 2. The best result was found when l λ = 2 with an average F-Measure of all folds of 0.8568. Figure 5 showed that the classification performance increase to a peak at λ = 2. Because the value of λ is a value that indicates the degree of importance of the occurrence of misclassification and the greater the value of λ, the smaller the error of classification that can be allowed.

Table 5. Effect of Parameter λ in degree 2 polynomial kernel

λ	Precision	Recall	F-Measure
0,01	0.8592	0.8379	0.8347
0,1	0.8592	0.8379	0.8347
1	0.8548	0.845	0.8422
2	**0.8766**	**0.859**	**0.8568**

The Experiments of γ Changes

Table 6 showed the results of γ changes to the SVM classifier performance using a degree 2 polynomial kernel. The experiments were carried out with λ = 2 and γ was started from 0.0001 which continues to 0.1. The best result was found at γ = 0.01 with an average F-Measure value of all folds of 0.8568. Figure 6 showed that the highest performance was at γ = 0.1. Experiments also showed that a learning rate that was too small gave poor results.

Table 6. Effect of γ Changes in Degree 2 Polynomial Kernel

γ	Precision	Recall	F-Measure
0,0001	0.6616	0.544	0.4223
0,001	0.8026	0.6839	0.6432
0,01	0.8796	0.845	0.8384
0,1	**0.8766**	**0.859**	**0.8568**

The SVM Model Selection for Degree 3 Polynomial Kernel

The Experiments of λ Changes

Table 7 showed the results λ changes to SVM classifier performance with a degree 3 polynomial kernel. The experiments was conducted with γ = 0.1 and λ was started from 0.01 which continued to 2. The best results were found when λ = 0.01 and λ = 0.1 with an average F-Measure value of all folds of 0.9919. As can be seen from Figure 7, the classification performance results were at their peak at = 0.01 and 0.01, after which they decreased. Because λ is the degree of importance of the occurrence of misclassification, then the greater the value of λ, the smaller the error of classification that can be allowed. In the classification model using a degree 3 polynomial kernel, overfitting started occured when λ = 1 so that the classification results drop drastically.

Table 7. Effect of λ Changes in Degree 3 Polynomial Kernel

λ	Precision	Recall	F-Measure
0,01	0.7831	0.715	0.6939
0,1	**0.7831**	**0.715**	**0.6939**
1	0.772	0.692	0.6673
2	0.752	0.693	0.6678

The Experiments of γ Changes

Table 8 showed the results of the effect of γ changes to the SVM classifier performance using a degree 3 polynomial kernel. The experiments were carried out with λ = 0.1 and γ was started from 0.0001 which continued to 0.1. The best result was shown when γ = 0.01 with the average F-Measure value of all folds of 0.75. As can be seen from Figure 7, that the classification performance result was at its peak at γ = 0.01 where after it has decreased. Experiments showed that a learning rate that is too small gives poor results.

Table 8. Effect of γ Changes in Degree 3 Polynomial Kernel

γ	Precision	Recall	F-Measure
0,0001	0.7127	0.548	0.4302
0,001	0.7571	0.6	0.5249
0,01	**0.8252**	**0.763**	**0.75**
0,1	0.7831	0.715	0.6939

The SVM Model Selection for RBF Kernel

The following are the results obtained from the experiments using the SVM with RBF kernel. Experiments were conducted via the k-fold cross validation method with k = 10.

The Experiments of λ Changes

Table 9 shows the results of the study on the effects of the λ changes to the SVM classifer performance using a RBF kernel. The experiments were carried out with γ = 0.1 and λ was started from 0.01 which continued to 2. The best result was found λ = 0.01 and 0.1 with an average F-measure value of all folds of 0.4494 (low value). From that experiments showed poor performance at learning rate (γ) = 0.1 .

Table 9. Effect of λ Changes in RBF Kernel

λ	Precision	Recall	F-Measure
0,01	0.716	0.5589	0.4494
0,1	**0.716**	**0.5589**	**0.4494**
1	0.7131	0.549	0.4316
2	0.6418	0.517	0.3711

The Experiments of γ Changes

Table 10 shows the results of the γ changes with to the SVM classifier performance with RBF kernel. The experiments were conducted with λ = 0.1 and γ was started from 0.0001 which continued to 0.1. The best result was found γ = 0.0001 and 0.001 with an average F-Measure value of all folds of 0.9676. From Figure 10 showed that the classification result was at its peak at γ = 0.001 where after it has decreased. The experiments showed that the greater the value of γ (learning rate), the worse the classification results.

Table 10. Effect of γ Changes in RBF Kernel

γ	Precision	Recall	F-Measure
0,0001	0.9702	0.968	0.9679
0,001	**0.9702**	**0.968**	**0.9679**
0,01	0.8547	0.794	0.7842
0,1	0.7160	0.5589	0.4494

Evaluation the Best Models

After conducting experiments to select the best model for each kernel (i.e., 5 models), the evaluation for measuring performance for those model using the testing data conducted. The best models obtained were the linear kernel model with λ = 0.1 and γ =0.1, the degree 1 polynomial kernel model with, λ = 0.1 and γ = 0.01, the degree 2 polynomial kenel model with λ = 2 and γ = 0.1, the degree 3 polynomial kernel model with λ = 0.1 and γ = 0.01, and RBF kernel model with λ = 0.1 and γ = 0.001

To find out unbiased results, it is necessary to use some testing data outside of the training data. The evaluation using the testing data showed that the linear kernel and the degree 1 polynomial kernel resulted the best performance of Precision, Recall, and F-Measure which were 0.99, 0.99, and 98%, respectively. It was compared to RBF which resulted Precision, Recall, and F-Measure which were 0.95, 0.95, and 0.94, respectively. The worst model was showed by the degree 3 polynomial kernel which returned Precision, Recall, and F1-Measure of 0.85, 0.79, and 0.78, respectively.

Table 11. Comparison of Results on SVM Kernel

Method	Precision	Recall	F-Measure
Linear	**0.9901**	**0.99**	**0.98**
Degree 1 Polynomial	**0.9901**	**0.99**	**0.98**
Degree 2 Polynomial	0.903	0.89	0.8891
Degree 3 Polynomial	0.8521	0.79	0.7803
RBF	0.9545	0.95	0.9487

Conclusion

Evaluation of the best model of each kernel model (i.e., 5 models: linear, degree 1, 2, 3 poynomial, and RBF kernel) using the testing data of identification Spam e-mail showed that the linear kernel and the kernel of degree 1 polynomial produced the best Precision, Recall, and F-Measure performances of 0.99, 0.99, and 98%, respectively. Compared to RBF kernel which produced Precision, Recall, and F-Measure of 0.95, 0.95, and 0.94, respectively. The worst model was shown by a degree 3 polynomial kernel which produces Precision, Recall, and F1-Measure of 0.85, 0.79, and 0.78, respectively.

Parameter λ in the SVM classification serves as the degree of importance of the occurrence of misclassification. The greater the value of λ, the greater the chance of overfitting and the smaller the value of λ, the greater the occurrence of underfitting. In this study, overfitting generally begins to occur at a value of λ = 1. Parameter γ in the SVM classification functions as a determinant of learning rate, where the greater the value, the F-Measure value of the SVM classification results will decrease. Learning rate that is too small gives poor results. In this study the results were low because the parameter value was too small, occurring at γ = 0.0001.

REFERENCE

[1] N. M. Samsudin, C. F. B. Mohd Foozy, N. Alias, P. Shamala, N. F. Othman, and W. I. S. Wan Din, “Youtube spam detection framework using naïve bayes and logistic regression,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 3, pp. 1508–1517, 2019, doi: 10.11591/ijeecs.v14.i3.pp1508-1517.

[2] K. Agarwal and T. Kumar, “Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization,” *Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018*, no. November, pp. 685–690, 2019, doi: 10.1109/ICCONS.2018.8662957.

[3] Y. Vernanda, S. Hansun, and M. B. Kristanda, “Indonesian language email spam detection using n-gram and naïve bayes algorithm,” *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 5, pp. 2012–2019, 2020, doi: 10.11591/eei.v9i5.2444.

[4] N. Sutta, Z. Liu, and X. Zhang, “A study of machine learning algorithms on email spam classification,” *EPiC Series in Computing*, vol. 69, pp. 170–179, 2020, doi: 10.29007/qshd.

[5] Z. Ge, “A fusion algorithm model based on KNN-SVM to classify and recognize spam,” *Journal of Physics: Conference Series*, vol. 1982, no. 1, 2021, doi: 10.1088/1742-6596/1982/1/012069.

[6] N. Sun, G. Lin, J. Qiu, and P. Rimba, “Near real-time twitter spam detection with machine learning techniques,” *International Journal of Computers and Applications*, vol. 44, no. 4, pp. 338–348, 2022, doi: 10.1080/1206212X.2020.1751387.

[7] S. N. D. Pratiwi and B. S. S. Ulama, “Klasifikasi Email Spam dengan Menggunakan Metode Support Vector Machine dan k-Nearest Neighbor,” *Jurnal Sains dan Seni ITS*, vol. 5, no. 2, pp. 344–349, 2016.

[8] A. R. Chrismanto and Y. Lukito, “Identifikasi Komentar Spam Pada Instagram,” *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 8, no. 3, p. 219, 2017, doi: 10.24843/lkjiti.2017.v08.i03.p08.

[9] Y. K. Zamil, S. A. Ali, and M. A. Naser, “Spam image email filtering using K-NN and SVM,”

International Journal of Electrical and Computer Engineering (IJECE), vol. 9, no. 1, p. 245, 2019, doi: 10.11591/ijece.v9i1.pp245-254.

[10] G. A. Reddy and B. I. Reddy, “Classification of Spam Text using SVM,” *Journal of University of Shanghai for Science and Technology*, vol. 23, no. 08, pp. 616–624, 2021, doi: 10.51201/jusst/21/08437.

[11] C. D. Manning, P. Raghavan, and S. Hinrich, *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2009. [Online]. Available: http://www.informationretrieval.org/

[12] R. J. Roiger, *Data Mining: A Tutorial-Based Primer*, 2nd ed. Boca Raton: CRC Press Taylor & Francis Group, 2017.

[13] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Boston: Pearson Addison-Wesley, 2006.

[14] C. C. Aggarwal, *Data Mining: The Textbook*. Switzerland: Springer, 2015. doi: 10.1007/978-3-319-14142-8.

[15] V. Jayaswal, “Performance Metrics: Confusion matrix, Precision, Recall, and F1 Score,” *Towards Data Science*, 2020. https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score-a8fe076a2262 (accessed Sep. 08, 2021).