 ISSN:2302-8149
e-ISSN:2540-9719

Perbandingan Pengujian Deteksi Phising menggunakan Metode SVM dengan Kernel *RBF* dan Linear

Comparison of Phishing Detection Tests using the SVM Method with RBF and Linear Kernels

Rumini*, Norhikmah, Ali Mustofa, Sulistyo Pradana
Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta
Jalan Ring Road Utara, Condong Catur, Depok, Sleman, Yogyakarta, Indonesia
*e-mail: rumini@amikom.ac.d

(***received*:** 14 Mei 2023, ***revised*:** 17 Juni 2023, ***accepted*:** 23 Juli 2023)

Abstrak

Phising adalah sebuah tindakan kriminal untuk mencuri informasi pribadi orang lain menggunakan entitas electronic, salah satunya adalah website. Informasi ini dicuri dari website yang telah diakses yang mengandung *phising* atau dengan kata lain masuk ke dalam kategori website phising. Tujuan dari web phising adalah membuat pengguna percaya bahwa mereka berinteraksi dengan situs resmi. Umumnya informasi yang dicari phisher (pelaku *phising)* adalah berupa *username, password*, baik itu akun media sosial atau akun nomor kartu kredit dengan cara diarahkan ke sebuah situs website palsu. Maka dari itu perlu adanya deteksi web phising yang bermanfaar untuk melindungi user dari tindak pencurian informasi pengguna. Penelitian ini membahas dua kernel dalam metode SVM (*Support Vector Machine)* untuk deteksi web *phising* yaitu kernel *RBF* (*Radial Basis Function)* dan *kernel linear*. Akurasi yang didapatkan dengan ketiga *kernel* menghasilkan nilai akurasi yang berbeda-beda. Hasil akurasi pengujian sistem deteksi web phising dengan Kernel Linear sebesar 92.582 % dan Kernel Radial Basis Function sebesar 96.426 %. Akurasi paling tinggi dengan metode SVM untuk deteksi web phising yaitu menggunakan kernel *RBF (Radial Basis Function)* yang dapat memberikan hasil deteksi yang akurat terhadap web yang mengandung *phising.*
Kata kunci: *SVM, phising, kernel, RBF, deteksi*

Abstract

Phishing is a criminal act to steal other people's personal information using electronic entities, one of which is a website. This information is stolen from websites that have been accessed which contain phishing or in other words enter into the category of phishing websites. The aim of web phishing is to make users believe that they are interacting with a legitimate site. Generally, the information that phishers are looking for is in the form of usernames, passwords, be it social media accounts or credit card account numbers by being directed to a fake website. Therefore, it is necessary to have phishing web detection which is useful to protect users from acts of theft of user information. This study discusses two kernels in the SVM (Support Vector Machine) method for web phishing detection, namely the RBF (Radial Basis Function) kernel and the linear kernel. The accuracy obtained with the three kernels produces different accuracy values. The results of the accuracy of testing the web phishing detection system with a Linear Kernel of 92,582% and a Radial Basis Function Kernel of 96,426%. The highest accuracy with the SVM method for detecting web phishing is using the RBF (Radial Basis Function) kernel which can provide accurate detection results for web containing phishing.
Keywords: *SVM, phising, kernel, RBF, detection*

1 Pendahuluan

Perkembangan Ilmu Pengetahuan dan Teknologi (IPTEK), khususnya pada bidang Teknologi dan Informasi saat ini sudah mengalami kemajuan yang pesat dan memberikan banyak manfaat, sehingga memudahkan hampir semua kegiatan penyebaran informasi, pengelolaan keuangan, dan metode pembayaran dapat dilakukan dengan Teknologi melalui jaringan internet dengan media

Website. Kemajuan Teknologi Informasi yang serba digital membawa orang berminat ke dalam dunia bisnis online karena dirasakan lebih mudah dan praktis berkomunikasi dan memperoleh informasi.

WeAreSocial merupakan sebuah perusahaan agensi marketing sosial terpercaya, yang mengeluarkan laporan setiap tahun mengenai data jumlah pengguna website, dan media sosial dari seluruh dunia. Laporan terbaru mengungkapkan bahwa pengguna internet mengalami perkembangan rata-rata lebih dari satu juta pengguna baru setiap hari. Jumlah orang yang menggunakan internet telah meningkat selama setahun terakhir. Ada 4,39 miliar pengguna internet pada tahun 2019, meningkat 366 juta(9 persen) dibandingkan Januari 2018. Rata-rata pengguna internet dunia menghabiskan 6 jam dan 42 menit online setiap hari[1]. Disisi lain, dengan kemajuan internet ini juga memiliki dampak negatif yang mengkhawatirkan pada berkembangnya tindak kejahatan online atau dapat disebut juga "CYBERCRIME". Disertai juga dengan maraknya website pishing.

Phising adalah sebuah tindakan kriminal untuk mencuri informasi pribadi orang lain menggunakan entitas electronic, salah satunya adalah website[2]. Informasi ini dicuri dari website yang telah diakses yang mengandung *phising* atau dengan kata lain masuk ke dalam kategori website *phising*. Tujuan dari web *phising* adalah membuat pengguna percaya bahwa mereka berinteraksi dengan situs resmi. Umumnya informasi yang dicari phisher(pelaku *phising*) adalah berupa username, password, baik itu akun media sosial atau akun nomor kartu kredit dengan cara diarahkan ke sebuah situs website palsu. Maka dari itu perlu adanya deteksi web *phising* yang berguna untuk melindungi user dari tindak pencurian informasi pengguna..

Menurut buku dengan judul No-tech hacking oleh Johnny Long mengutarakan bahwa senjata paling ampuh dalam dunia hacker adalah social engineering[3]. Serangan Web *Phising* memuncak pada tahun 2005 dan *phising* merupakan cara untuk memikat agar korban dengan mudah jatuh dalam perangkap penipuan, dengan mencari kelemahan di dalam website ,dan kelemahan e-mail, pada dasarnya *phising* menggunakan hampir semua teknik hacking yang digunakan untuk membuat umpan[4. Hacking merupakan tindak kejahatan komputer untuk mengakses, mencuri, merusak data secara tidak sah[5].

Support Vector Machine (SVM) dikembangkan oleh Bose, Guyon, dan Vapnik, pertama kali diperkenalkan pada tahun 1992 di Annual Workshop On Computational Learning Theory. *Support Vector Machine* merupakan salah satu metode dalam supervised learning yang biasanya digunakan untuk klasifikasi seperti Support Vector Classification dan regresi Support Vector Regression. SVM juga dapat mengatasi masalah klasifikasi dan regresi linear maupun non linear[6].

Tujuan dalam penelitian ini adalah untuk menerapkan Algoritma *Support Vector Machine* (SVM) untuk klasifikasi Web *Phising* untuk mengklasifikasikan dan memaksimalkan Data Set yang akan diolah, Sehingga dapat memberikan manfaat dalam menghasilkan akurasi yang maksimal untuk mendeteksi web *phising*. Dengan menerapkan hasil akurasi tertinggi dari 2 kernel yang diuji yaitu kernel linear dan kernel *RBF*

2 Tinjauan Literatur

Tinjauan Literatur mengenai penelitian-penelitian yang terkait. Penelitian sebelumnya dengan judul "Prediksi Website Pemancing Informasi Penting *Phising* Menggunakan *Support Vector Machine* (SVM)" Algoritma yang digunakan *Support Vector Machine*, Topik yang diteliti sama yaitu web *phising*, Menggunakan dataset publik yang sama yaitu dari *UCI Machine Learning Repository* Sedangkan perbedaan antara keduanya terletak pada penerapan dan perbandingan yang dilakukan oleh Zuhri Halim, pada penelitian yang dilakukan oleh Zuhri Halim adalah melakukan perbandingan antara *Support Vector Machine* dengan *Naive Bayes* dan *Decision Tree*. Sedangkan dalam penelitian yang saya lakukan adalah untuk menerapkan Algoritma *Support Vector Machine* untuk klasifikasi web *phising*. Dan berbeda pada bagian *pre-processing* [7].

Peelitian sebelumnya dengan judul "Klasifikasi Email Spam dengan Menggunakan Metode *Support Vector Machine* dan *k-Nearest Neighbor"* Algoritma yang digunakan Support Vector Machine. Topik yang diteliti sama yaitu web *phising*. Sedangkan perbedaan antara keduanya terletak pada sumber data yang digunakan oleh Sheila Novelia dan Brodjol Sutijo , yaitu menggunakan 6000 email dengan format hml. Sedangkan dataset yang saya gunakan adalah *UCI Machine Learning Repository* (www.uci.edu). Pada penelitian yang dilakukan oleh Sheila Novelia dan Brodjol Sutijo adalah melakukan perbandingan antara *Support Vector Machine* dengan *k-Nearest Neighbor*. Sedangkan

 ISSN:2302-8149
e-ISSN:2540-9719

dalam penelitian yang saya lakukan adalah untuk menerapkan Algoritma *Support Vector Machine* untuk klasifikasi web *phising*. Dan berbeda pada bagian *pre-processing*. Dan terdapat perbedaan juga pada object yang diteliti yaitu pada penelitian Sheila Novelia dan Brodjol Sutijo menggunakan *object* email untuk diklasifikasi, Sedangkan dalam penelitian saya menggunakan *object* web *phising* untuk diklasifikasi [8].
Pada Judul, "Phishing Websites Classification using Hybrid *SVM* and *KNN* Approach", Algoritma yang digunakan Support Vector Machine. Sedangkan perbedaan antara keduanya terletak pada sumber data, sedangkan dataset yang saya gunakan adalah *UCI Machine Learning Repository* (www.uci.edu). Pada penelitian yang dilakukan oleh Altyeb Altaher adalah mengintegrasi antara *Support Vector Machine* dengan *k-Nearest Neighbor*. Sedangkan dalam penelitian yang saya lakukan adalah untuk menerapkan Algoritma *Support Vector Machine* untuk klasifikasi web *phising* [9].

3 Metode Penelitian

Metode penelitian menggunaka Metode *Support Vector Machine* (SVM). Alur penelitian adalah melakukan perencanaan dan pengumpulan data, data Web *Phising* dan *non Phising* didapatkan dari UCI *Machine Learning Repository*. Proses training set merupakan bagian dataset latih untuk membuat prediksi atau menjalankan fungsi dari sebuah algoritma. Hasil dari training set akan menjadi acuan bagi sistem untuk menentukan sebuah inputan data web *phising* atau bukan *phishing*, untuk melihat keakuratan/performanya. Proporsi test set sebesar 20% dan train set 80%.

Selanjutnya dilakukan proses perhitungan SVM, yaitu klasifikasi dengan kernel *linear* dan kernel *gaussian* (*radial basis function*, *RBF*) dengan melakukan import library SVM dengan query form sklearn import svm, dan meng-import library *accuracy_score* dari library from *sklearn.metrics*. Library SVM berfungsi untuk melakukan klasifikasi dengan menggunakan SVC(*Support Vector Clasification*), SVC merupakan perintah untuk menjalankan model SVM. Library *accuracy_score* memiliki fungsi untuk menampilkan hasil dari klasifikasi yang dilakukan oleh SVM. Selanjutnya dilakukan pengujian dalam proses klasifikasinya.

Fungsi kernel yang paling sederhana adalah kernel linear yang digunakan untuk klasifikasi data linear, kernel linear digunakan ketika data yang dianalisis sudah terpisah secara linear. Kernel linear cocok ketika terdapat banyak fitur dikarenakan pemetaan ke ruang dimensi yang lebih tinggi tidak benar-benar meningkatkan kinerja. Persamaan untuk fungsi kernel linear adalah:

$$\boldsymbol{K(x, x_k) = x_k^t x} \quad (1)$$

Sedangkan untuk *kernel RBF* (*Radial basis fucnctuon/Gaussian*) digunakan untuk klasifikasi data non-linear, *Kernel RBF* atau juga disebut *kernel Gaussian* adalah konsep kernel yang paling banyak digunakan untuk memecahkan masalah klasifikasi data yang tidak dapat dipisahkan secara linear. Kernel ini dikenal memiliki performa yang baik dengan parameter tertentu, dan hasil dari pelatihan memiliki nilai error yang kecil dibandingkan dengan kernel lainnya. Rumus persamaan untuk fungsi kernel *RBF* adalah [10]:

$$\boldsymbol{K(x, x_k) = exp\{-||x - x_k\}||_2^2/\sigma^2} \quad (2)$$

Hasil akurasi yang didapatkan dengan melakukan pengujian deteksi *phising* dengan metode SVM *kernel RBF* dan *kernel Linear*. Hasil pengujian dengan akurasi terbaik untuk deteksi *phising* adalah dengan SVM kernel *RBF*. Penarikan kesimpulan dirumuskan berdasarkan analisis hasil pengujian dan mengacu pada tujuan dari penelitian yang dilakukan.

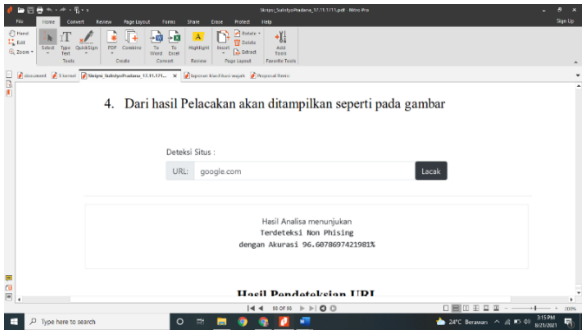
4 Hasil dan Pembahasan

Dalam penelitian ini penulis melakukan pengujian dengan 2 cara. Pertama adalah dengan pengujian melalui perbandingan jumlah data training dan testing, yang bertujuan untuk mencari nilai akurasi terbesar dengan proporsi *test set* sebesar 20% dan *train set* 80%. Kedua adalah pengujian jenis kernel, bertujuan untuk mengetahui rata-rata akurasi terbaik dari kernel yang akan digunakan.

 ISSN:2302-8149
e-ISSN:2540-9719

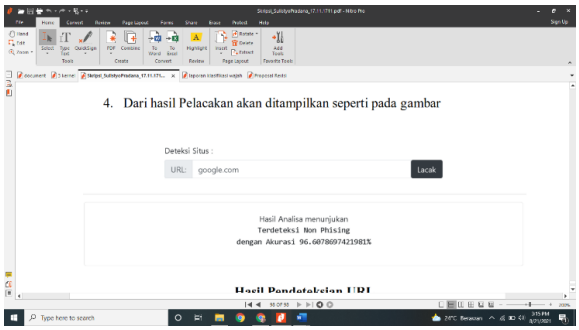
3.1. Pengujian Deteksi *Phising*

Pengujian deteksi *phising* dengan langkah-langkah pengujian deteksi *phising* yaitu menyiapkan URL yang dicurigai sebagai phising, sebagai contoh, https://tallals.com/, https://kakudobno.ru, https://recambiosscooter.es/signin/ dan menyiapkan URL yang tidak dicurigai sebagai phising, contoh, https://google.com/, https://dicoding.com/ . Gambar 2 menunjukkan hasil deteksi sistem untuk klasifikasi web *non- phising* yaitu situs google dengan akurasi 96,68%.



Gambar 1. Hasil deteksi "*non-Phising*"

Pada tahapan uji klasifikasi dari gambar 3 menunjukkan bahwa situs https://kakudobno.ru merupakan situs yang terdeteksi *web phising* dengan nilai akurasi sebesar 96,60%.



Gambar 2. Hasil deteksi "*phising*"

3.2. Pengujian Jenis Kernel

Pada penelitian ini dilakukan perbandingan hasil akurasi dari klasifikasi *Support Vector Machine* menggunakan dua kernel yang berbeda yaitu *kernel Linear* dan *kernel Radial Basis Function (RBF).* Berikut hasil klasifikasi *Support Vector Machine* dengan data test dan data train sebanyak 11056 dari UCI Machine Learning. Hasilnya ada pada Tabel 1.

Tabel 1. Hasil Pengujian *Kernel Linear* Dan *Rbf*

ALGORITMA	KERNEL	HASIL AKURASI
SVM	Linear	92.582 %
SVM	*RBF*	96.426%

3.1 *Confusion Matrix*

Confusion matrix suatu metode yang biasanya digunakan untuk melakukan perhitungan akurasi pada konsep data mining atau Sistem Pendukung Keputusan. Pada pengukuran kinerja menggunakan *confusion matrix*, terdapat 4 (empat) istilah sebagai representasi hasil proses klasifikasi. Keempat istilah tersebut adalah *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) dan *False Negative* (FN). Nilai *True Negative* (TN) merupakan jumlah data negatif yang terdeteksi

 ISSN:2302-8149
e-ISSN:2540-9719

dengan benar, sedangkan *False Positive* (FP) merupakan data negatif namun terdeteksi sebagai data positif. Sementara itu, *True Positive* (TP) merupakan data positif yang terdeteksi benar. *False Negative* (FN) merupakan kebalikan dari *True Positive*, sehingga data posifit, namun terdeteksi sebagai data negatif. Pada tabel menunjukan hasil dari confusion matrix metode *Support Vector Machine*.

Tabel 2. Hasil *Confution Matrix* Metode *Support Vector Machine*

	True no Phising	*True Phising*
Pred. No phising	911	52
Pred phising	35	1213

Jumlah True Positive (TP) adalah 911 diklasifikasikan sebagai 1 sesuai dengan prediksi yang dilakukan dengan Algoritma Support Vector Machine, lalu False Negative (FN) sebanyak 52 data diprediksi sebagai 1 tetapi ternyata -1, kemudian True Positive (TN) sebanyak 1213 data sesuai dengan prediksi,kemudian False Positife (FP) sebanyak 35 data diprediksi -1 ternyata 1. Tingkat akurasi yang dihasilkan dengan menggunakan Algoritma Support Vector Machine adalah sebesar 96% dan dapat dihitung untuk mencari nilai *accuracy.*

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{911+1213}{911+1213+35+52} = 0{,}9606 \quad (1)$$

Keterangan =

TP = True Positive
TN = True Negative
FN= False Negative
FP= Fasle Positive

3.3 Evaluasi Pengujian

Klasifikasi web *Phising* dengan menggunakan Algoritma Support Vector Machine dengan *Kernel Linear* dan *Kernel RBF* terhadap *data test* dan *data training* telah selesai dilakukan. Dari hasil pengujian dengan menggunakan data test dan data training yang sama dengan banyaknya dataset 11056 didapatkan hasil akurasi pengujian dengan Kernel Linear sebesar 92.582 % dan *Kernel Radial Basis Function* sebesar 96.426 %. *Algoritma Support Vector Machine* dengan *kernel radial basis function* untuk klasifikasi Website *Phising* memiliki akurasi yang lebih baik daripada *kernel linear*. Dari pengujian yang dilakukan didapatkan nilai akurasi sebesar 96.426 %. Dari hasil pengujian tersebut dapat dilihat bahwa *Algoritma Support Vector Machine* dapat digunakan untuk mengklasifikasikan Website *Phising*.

5 Kesimpulan

Berdasarkan hasil penelitian bahwa Algoritma *Support Vector Machine* telah berhasil diterapkan untuk deteksi *Phising* dan mendapatkan hasil akurasi yang lebih baik daripada penelitian sebelumnya, berdasarkan hasil pengujian kernel dapat diketahui bahwa hasil akurasi kernel *Radial Basis Function* (*RBF*) lebih baik daripada kernel Linear untuk mengklasifikasikan Web Phising. Hasil akurasi pengujian dengan Kernel *Linear* sebesar 92.582 % dan Kernel *Radial Basis Function* (*RBF*) sebesar 96.426 %

Referensi

[1] "Data," 2022. https://databoks.katadata.co.id/datapublish/2022/03/23/ada-2047-juta-pengguna-internet-di-indonesia-awal-2022
[2] R. M. Mohammad, F. Thabtah, and L. McCluskey, "An assessment of features related to phishing websites using an automated technique," in *2012 International Conference for Internet Technology and Secured Transactions, ICITST 2012*, 2012, pp. 492–497.

 ISSN:2302-8149
e-ISSN:2540-9719

[3] B. S. Dharma Pratiwi, Sheila Noveila, Suprih Ulama, "Klasifikasi Email Spam dengan Menggunakan Metode Support Vector Machine dan k-Nearest Neighbor," *J. Sains Dan Seni Its Vol. 5No.2(2016)2337-3520(2301-928XPrint)*, vol. 5, 2016, [Online]. Available: http://ejurnal.its.ac.id/index.php/sains_seni/article/view/16685
[4] Weiss S., *Text Mining : Perspective Methods for Analysis and Prediction Model Discovery Using RapidMiner*. Indurkhya, edutor. New Jersey L Springer Science & Business Media., 2010.
[5] J. S. R. Feldman, *The Text Mining Handbook*. New York: Cambridge University Press, 2007.
[6] S. B. Maryuni, "Identifikasi Website Phising Dengan Seleksi Atribut Berbasis Korelasi," *Semin. Nas. Teknol. dan Komunikasi(SENTIKA)*.
[7] Z. Halim, "Prediksi Website Pemancing Informasi Penting Phising Menggunakan Support Vector Machine (SVM)," *Inf. Syst. Educ. Prof.*, vol. 2, no. 1, pp. 71–82, 2017, [Online]. Available: http://download.portalgaruda.org/article.php?article=535068&val=10928&title=Prediksi Website Pemancing Informasi Penting Phising Menggunakan Support Vector Machine (SVM)
[8] S. N. D. Pratiwi and B. S. S. Ulama, "Klasifikasi Email Spam dengan Menggunakan Metode Support Vector Machine dan k-Nearest Neighbor," *J. Sains dan Seni ITS*, vol. 5, no. 2, pp. 344–349, 2016.
[9] A. Altaher, "Phishing Websites Classification using Hybrid SVM and KNN Approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 6, pp. 90–95, 2017, doi: 10.14569/ijacsa.2017.080611.
[10] Suyanto, *Machine Learning, Tingkat dasar dan Lanjut*. Bandung:Informatika, 2018.
[11] "'What is Python Good For?'. General Python FAQ. Python Software Foundation." https://docs.python.org/3/faq/general.html#what-is-python-good-for.
[12] "'What is Python? Executive Summary'. Python documentation. Python Software Foundation." https://www.python.org/doc/essays/blurb/
[13] "General Python FAQ," *python.org. Python Software Foundation.* https://docs.python.org/3/faq/general.html#what-is-python.
[14] and F. T. Abdelhamid, Neda, Aladdin Ayesh, *Phishing detection based associative classification data mining.‖ Expert Systems with Applications*. 2014.