PERBANDINGAN KINERJA ALGORITMA NAÏVE BAYES DAN C4.5 UNTUK MENDETEKSI PENGELABUAN UNIFORM RESOURCE LOCATOR (PHISHING URL)

Kevin Marcello Jonathan [1)] Bagus Mulyawan [2)] Novario Jaya Perdana [3)]

[1), 2), 3)] Teknik Informatika, FTI, Universitas Tarumanagara
Jl. Letjen S Parman No.1, Jakarta 11440 Indonesia
Email: kevin.535160019@stu.untar.ac.id[1)], bagus@fti.untar.ac.id[2)] novariojp@fti.untar.ac.id[3)]

ABSTRACT

Nowadays, phone or smartphone and internet are something that cannot be separated with human. One of negative effects of using internet is cyber crime like a phishing url. Phishing url is usually used to collecting personal information like pin number, credit card etc. There are many type of classification algorithm, two of them is Naïve Bayes and C4.5. Both of the alogithm is good for recognizing a phishing url. This website created are used to classify a unkown url with Naïve Bayes and C4.5 algorithm. The accuracy of C4.5 algorithm is 87.11% and 78.48% for Naïve Bayes. The average time needed to processing one url is 21.78 second for C4.5 and 23.31 second for Naïve Bayes.

Key words

Naïve Bayes, C4.5, Uniform Resource Locator, Perbandingan, Phishing

1. Pendahuluan

Semakin berkembangnya teknologi, kejahatan rentan terjadi dimana saja, termasuk melalui Uniform Resource Locator atau yang biasa disingkat URL. URL adalah rangkaian karakter menurut suatu format standar tertentu, yang digunakan untuk menunjukkan alamat suatu sumber seperti dokumen dan gambar di Internet[1]. URL digunakan untuk mengidentifikasi lokasi sebuah file dalam internet. URL digunakan tak hanya untuk membuka sebuah situs web, tetapi juga untuk mengunduh video, gambar, halaman hypertext, dan yang lainnya.

Beberapa jenis kejahatan yang ditemui dalam sebuah url salah satunya yaitu phishing. Phishing adalah singkatan dari password harvesting fishing adalah tindakan penipuan yang menggunakan email palsu atau situs website palsu yang bertujuan untuk mengelabui user sehingga pelaku dapat mendapatkan data user tersebut [2]. Tindakan penipuan ini berupa sebuah email yang seolah-olah berasal dari sebuah perusahaan resmi, misalnya bank dengan tujuan untuk mendapatkan data-data pribadi seseorang, misalnya PIN, nomor rekening, nomor kartu kredit, dan sebagainya.

Phishing dapat dideteksi dengan beberapa metode,ssalah satunya adalah Naïve Bayes. Naïve Bayes sebuah metode klasifikasi yang menggunakan metode probabilitas dan statistik untuk memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya. Ciri utama dari Naïve Bayes ini adalah asumsi yang sangat kuat (naïf) akan independensi dari masing-masing kondisi / kejadian.

Algoritma C4.5 merupakan kelompok algoritma Decision Tree. Algoritma ini mempunyai input berupa training samples dan samples. Training samples berupa data contoh yang akan digunakan untuk membangun sebuah tree yang telah diuji kebenarannya. Sedangkan samples merupakan field-field data yang nantinya akan digunakan sebagai parameter dalam melakukan klasifikasi data.

Pada penulisan makalah ini akan dibahas mengenai website berbasis PHP untuk melakukan pengenalan terhadap url asing dengan menggunakan algoritma Naïve Bayes dan C4.5 sebagai algoritma pengklasifikasian. Masukan perangkat lunak ini berupa citra alamat url asing yang ingin dideteksi serta keluarannya merupakan hasil perhitungan klasifikasi dari url.

Tahapan pembuatan dimulai dari pengumpulan url dari berbagai sumber yang bersifat phising maupun bersifat aman. Setelah dicari, maka url akan dimasukan kedalam database dengan terlebih dahulu di ekstrak dengan mengunakan beberapa fitur url [4][5]. Setelah data dikumpulkan, dilakukan perhitungan dengan menggunakan kedua algoritma, dan kemudian akan dicari besaran akurasi serta lama proses perhitungan dari masing – masing algoritma.

2. Dasar Teori

2.1Phishing

Menurut IGN Mantra, phishing adalah percobaan penipuan menggunakan surel (surat elektronik) dengan tujuan untuk mendapatkan username, password, token, dan informasi–informasi sensitive lainnya yang dikirim melalui surel. Surel phishing datang seolah-olah berasal dari perusahaan/organisasi di mana user adalah anggota/member [6].

Adapun untuk mendapatkan korban phishing, banyak

cara yang digunakan dan hal ini biasanya terus berkembang sesuai dengan perkembangan yang ada di dalam dunia internet. Beberapa metode yang populer digunakan adalah [7].

1. Email / SPAM
   Media yang paling favorit digunakan untuk mencari korban adalah email. Email dipilih karena murah dan mudah untuk digunakan. Pelaku dapat mengirimkan jutaan email setiap harinya tanpa perlu mengeluarkan biaya yang cukup besar. Bahkan pelaku phishing juga suka menggunakan server-server bajakan untuk melakukan aksinya.
2. Web-based Delivery
   Pelaku phishing juga memanfaatkan website dalam melakukan aksinya. Pelaku biasanya membuat website yang mirip dengan website-website terkenal untuk mengelabui korbannya.Membuat website yang mirip dengan website perusahaan besar sangatlah mudah untuk dilakukan karena pelaku hanya perlu membuat tampilan yang sama, tanpa perlu membuat fungsi atau fasilitas yang sama karena tujuannya adalah agar korban memasukkan username dan password di dalamnya kemudian korban akan dibawa ke situs asli agar tidak curiga.
3. IRC / Instant Messaging
   Media chatting yang banyak digunakan sebagai sarana pelaku phishing untuk mengirimkan alamat-alamat yang menjebak kepada korbannya. Biasanya pelaku mengirimkan link ini secara acak namun ada juga yang melakukan pendekatan terlebih dahulu sebelum mengirimkan informasi situs palsu ini.
4. Trojan
   Pelaku phishing, terkadang juga menipu korbannya agar menginstall trojan dan memanfaatkan trojan tersebut untuk mengelabui korbannya. Trojan memungkinkan pengontrolan secara penuh komputer korban sehingga korban dapat dialihkan ke situs yang telah disediakan jebakan.

2.2Data Mining

Data Mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait berbagai basis data besar [8].

Didalam data mining, data yang disimpan secara elektronik dapat diolah secara otomatis oleh komputer. Pengolahan data yang dilakukan dapat menghasilkan suatu pola tertentu, mengidentifikasi, validasi, dan prediksi dari suatu data secara otomatis dalam berbagai bidang, seperi ekonomi, statistik, ramalan cuaca, dan bidang lainnya [9]. Data Mining secara operasional adalah proses menemukan pola secara otomatis atau semi-otomatis dalam sejumlah data yang besar dan pola tersebut harus menguntungkan.

2.3Algoritma Naïve Bayes

Bayes merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema Bayes (atau aturan Bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat (naïf). Dengan kata lain, Naïve Bayes, model yang digunakan adalah "model fitur independen" [10].

Dalam Bayes (terutama Naïve Bayes), maksud independensi yang kuat pada fitur adalah bahwa sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama.

Prediksi Bayes didasarkan pada teorema Bayes dengan formula umum sebagai berikut [10]:

$$P(C|X) = \frac{P(X|C)\,P(c)}{P(x)} \quad \text{.....(1)}$$

Keterangan:
x = data yang belum diketahui
c = hipotesis data merupakan suatu kelas spesifik
P(c|x) = Probabilitas hipotesis bedasarkan kondisi (posteriori probability)
P(c) = Probabilitas hipotesis (prior probability)
P(x|c) = Probabilitas bedasarkan kondisi pada hipotesis
P(x) = Probabilitas c

Ide dasar dari aturan Bayes adalah bahwa hasil dari hipotesis atau peristiwa dapat diperkirakan berdasarkan pada beberapa bukti yang diamati. Ada beberapa hal penting dari aturan Bayes tersebut, yaitu :

1. Sebuah probabilitas awal/prior H atau P(H) adalah probabilitas dari suatu hipotesis sebelum bukti diamati.
2. Sebuah probabilitas akhir H atau P(H|E) adalah probabilitas dari suatu hipotesis setelah bukti diamati.

Rumus di atas menjelaskan bahwa peluang bahwa peluang masuknya sampel karakteristik tertentu dalam kelas C (Posterior) adalah peluang munculnya kelas C (sebelum masuknya sampel tersebut, seringkali disebut prior), dikali dengan peluang kemunculan karakteristik karakteristik sampel pada kelas C (disebut juga likelihood), dibagi dengan peluang kemunculan karakteristik sampel secara global (disebut juga evidence).

2.4 Algoritma C4.5

C4.5 adalah sebuah algoritma yang digunakan untuk memproduksi sebuah decision tree yang merupakan ekspansi dari pendahulunya yaitu kalkulasi ID3 [11]. Algoritma ini meningkatkan algoritma ID3 dengan mengatur properti yang berkelanjutan dan berlainan, missing value dan pemotongan tree setelah konstruksi. Decision tree yang dibuat dengan C4.5 dapat digunakan untuk pengelompokkan dan seringkali cenderung mengarah ke statistical classifier. C4.5 membuat decision tree dari sebuah set data sampel sama seperti algoritma ID3. Sebagai algoritma pembelajaran yang terkelola, algoritma ini membutuhkan sebuah set dari data sampel yang dapat terlihat seperti data pasangan : objek masukan dan nilai keluaran yang diinginkan (class).

Pemilihan atribut yang baik adalah atribut yang memungkinkan untuk mendapatkan decision tree yang

paling kecil ukurannya, atau atribut yang dapat memisahkan obyek menurut kelasnya. Secara heuristik atribut yang dipilih adalah atribut yang menghasilkan simpul yang paling purest (paling bersih). Ukuran purity dinyatakan dengan tingkat impurity, dan untuk menghitungnya, dapat dilakukan dengan menggunakan konsep Entropy, Entropy menyatakan impurity suatu kumpulan objek .

Formula mencari entropi sebagai berikut [11] :

$$Entropi\ (S) = \sum_{j=1}^{k} - p_{j\ log_2\ p_j} \text{.....(2)}$$

Keterangan:

S = Himpunan (dataset) kasus
K = Banyaknya partisi S
Pj = Probabilitas yang di dapat dari Sum(Ya) dibagi total kasus

Information gain adalah kriteria yang paling populer untuk pemilihan atribut. Algoritma C4.5 adalah pengembangan dari algoritma ID3. Oleh karena pengembangan tersebut algoritma C4.5 mempunyai prinsip dasar kerja yang sama dengan algoritma ID3. Hanya saja dalam algoritma C4.5 pemilihan atribut dilakukan dengan menggunakan Gain Ratio dengan rumus [11] :

$$gain\ ratio\ (a) = \frac{gain(a)}{split(a)} \text{.....(3)}$$

Keterangan :

a = atribut
gain(a) = information gain pada atribut a
Split(a) = split information pada atribut a

Atribut dengan nilai Gain Ratio tertinggi dipilih sebagai atribut test untuk simpul. Dengan gain adalah information gain. Pendekatan ini menerapkan normalisasi pada information gain dengan menggunakan apa yang disebut sebagai split information. SplitInfo menyatakan entropy atau informasi potensial dengan rumus [11] :

$$SplitInfo(S, A) = -\sum_{i-1}^{n} \frac{S_i}{S} \log_2 \frac{Si}{s} \text{.....(4)}$$

Keterangan :
S = ruang (data) sample yang digunakan untuk training.
A = atribut.
Si = jumlah sample untuk atribut i

Sedangkan untuk mencari nilai Gain digunakan rumus [11] :

$$Gain(A) = Entropi(S) \sum_{i=1}^{k} \frac{|Si|}{|S|} xEntropi(Si) \text{.....(5)}$$

Keterangan :
S = ruang (data) sample yang digunakan untuk training.
A = atribut.
|Si| = jumlah sample untuk nilai V.
|S| = jumlah seluruh sample data.
Entropi(Si) = entropy untuk sample-sample yang memiliki nilai i

2.5 Confussion Matrix

Pengukuran terhadap kinerja suatu sistem klasifikasi merupakan hal yang penting. Kinerja sistem klasifikasi menggambarkan seberapa baik sistem dalam mengklasifikasikan data. Confusion matrix merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi. Pada dasarnya confusion matrix mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya [12].

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 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 positif, namun terdeteksi sebagai data negatif.

Berdasarkan nilai True Negative (TN), False Positive (FP), False Negative (FN), dan True Positive (TP) dapat diperoleh nilai akurasi, presisi dan recall. Nilai akurasi menggambarkan seberapa akurat sistem dapat mengklasifikasikan data secara benar. Dengan kata lain, nilai akurasi merupakan perbandingan antara data yang terklasifikasi benar dengan keseluruhan data.

Berikut persamaan untuk mencari akurasi [12]:

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \text{.....(6)}$$

Keterangan :
TP = True Positive, yaitu jumlah data positif yang terklasifikasi dengan benar oleh sistem.
TN = True Negative, yaitu jumlah data negatif yang terklasifikasi dengan benar oleh sistem.
FN = False Negative, yaitu jumlah data negatif namun terklasifikasi salah oleh sistem.
FP = False Positive, yaitu jumlah data positif namun terklasifikasi salah oleh sistem

Tabel 1 Hasil Gambaran Confussion Matrix

		Aktual	
		Positif	Negatif
Prediksi	Positif	True Positive	False Positive
	Negatif	False Negative	True Negatif

3. Hasil Percobaan dan Pembahasan

Pengujian dilakukan dengan menggunakan total 107225 alamat url yang didapatkan dari berbagai sumber [13][14][15]. Setelah url didapatkan maka langkah berikutnya adalah mengekstraksi fitur – fitur dari url yang kemudian akan dimasukan kedalam database beserta jenis url. Disini url akan diberi label "0" apabila bukan url phishing dan diberi label "1" apabila url tersebut merupakan url phishing.

Fitur – fiitur yang digunakan untuk mengekstraksi alamat url Fitur pertama url berupa IP address. Jika url berupa ip address maka akan diberi nilai 1, jika tidak akan diberi nilai 0 Fitur kedua panjang url kurang dari lima puluh empat atau lebih dari 75. Jika kurang dari lima uluh empat maka akan diberi nilai -1, bila panjang url diantara 54 sampai 75 maka akan diberi nilai 0 dan jika panjang url lebih dari 75 maka akan diberi nilai 1

Fitur ketiga url memiliki symbol "@". Jika alamat url memiliki symbol "@" maka akan diberi nilai 1, jika tidak mengandung symbol "@" maka akan diberi nilai Fitur keempat url memiliki symbol "-". Jika alamat url memiliki symbol "-" maka akan diberi nilai 1, jika tidak mengandung symbol "-" maka akan diberi nilai 0. Fitur kelima url memiliki "." pada subdomain. Jika subdomain memiliki "." kurang dari dua maka akan diberi nilai -1, jika "." sama dengan dua maka akan diberi nilai 0 dan bila "." lebih dari dua maka akan diberi nilai 1.

Fitur keenam alamat url menggunakan "https" dan kejelasan alamat url terpecaya. Fitur ini melihat apakah ijin "https" pada alamat url memiliki ijin dari penerbit yang jelas serta diukur dari lamanya usia dari sertifikat yang dimiliki alamat url. Apabila alamat url mengandung dan memiliki sertifikat "https" serta berumur diatas satu tahun, maka url akan diberi nilai -1. Apabila alamat url mengandung https namun tidak tertera pada list penerbit yang jelas ('Comodo' , 'Symantec' , 'GoDaddy' , 'GlobalSign' , 'DigiCert' , 'StartCom' , 'Entrust' , 'Verizon' , 'Trustwave' , 'Unizeto' , 'Buypass' , 'QuoVadis' , 'Deutsche Telekom' , 'Network Solutions' , 'SwissSign' , 'IdenTrust' , 'Secom' , 'TWCA' , 'GeoTrust' , 'Thawte' , 'Doster' , 'VeriSign') maka akan diberi nilai 0. Jika alamat url tidak memiliki fitur seperti diatas maka akan diberi nilai 1.

Fitur ketujuh domain alamat url berusia lebih dari satu tahun atau tidak. Apabila alamat url masa berlaku domainnya kurang dari satu tahun maka akan diberi nilai 1 dan jika masa berlaku domain lebih dari satu tahun maka akan diberi nilai 0. Fitur kedelapan alamat url mengandung "htttps" pada alamat domain url. Apabila domain pada url mengandung kata "https" maka akan diberi nilai 1 dan jika tidak mengandung maka akan diberi nilai 0. Fitur kesembilan seberapa banyak alamat url meminta tautan dari url lain, seperti gambar dan video maupun suara yang diambil dari domain url lain. Apabila jumlah media atau objek yang diambil dari domain lain berjumlah kurang dari dua puluh dua persen maka akan diberi nilai -1, jika berada diantara dua puluh dua sampai enam puluh satu persen maka akan diberi nilai 0 dan jumlahnya melebihi enam puluh satu persen maka akan diberi nilai 1.

Fitur kesepuluh menghitung banyaknya anchor "<a>" pada alamat url. Fitur ini kurang lebih mirip dengan fitur kesembilan, namun memiliki beberapa perbedaan. Ada dua syarat yang menjadikan "<a>" dianggap mencurigakan yaitu : bila "<a>" dan alamat url memiliki domain yang berbeda, serta "<a>" tidak merujuk ke alamat domain apapun seperti (<a href="#">, <a href="#content">, <a href="#skip">, <a href ="Javascript::void(0)">). Apabila jumlah "<a>" yang terkandung pada url dan memenuhi dua syarat diatas jumlahnya dibawah tiga puluh satu persen didalam url maka akan diberi nilai -1, jika terkandung diantara tiga puluh satu sampai enam puluh tujuh persen akan diberi nilai 0 dan jika melebihi enam puluh tujuh persen maka akan diberi nilai 1.

Fitur kesebelas menghitung banyaknya tag didalam "<meta>, <script> dan <links>". Apabila jumlah link yang terkandung pada ketiga bagian tersebut kurang dari tujuh belas persen maka akan diberi nilai -1, jika terkandung diantara tujuh belas sampai delapan puluh satu akan diberi nilai 0 dan jika melebihi delapan puluh satu persen maka akan diberi nilai 1. Fitur kedua belas mencari fungsi memasukan informasi pribadi ke suatu alamat email tertentu ('mail()' atau 'mailto:'). Jika ditemukan fungsi tersebut didalam alamat url maka akan diberi nilai 1, jika tidak akan diberi nilai 0. Fitur ketiga belas umur dari domain diatas enam bulan atau tidak dengan menggunakan database dari fitur WHOIS. Jika umur domain diatas enam bulan maka akan diberi nilai 0 dan apabila umur domain dibawah enam bulan maka akan diberi nilai -1.

Fitur keempat belas sampai kedelapan bela url mengandung tulisan atau nama brand (confirm, log in, sign in, eBay, PayPal). Jika url mengandung tulisan atau nama brand diatas akan diberikan nilai 1 dan jika tidak maka akan diberi nilai 0.

3.1Hasil Pengujian Algoritma

Pengujian dilakukan dengan membagi database menjadi empat proporsi pembagian, yaitu 65% training – 35% testing, 70% training – 30% testing, 75% training – 25% testing dan 80% training - 20% testing. Data yang digunakan sebagai pembanding yaitu berjumlah 100 url. Untuk database yang berjumlah 107225 juga dilakukan perhitungan. Untuk menguji kecepatan dalam melakukan perhitungan pada suatu alamat url, digunakan 10 alamat url untuk diketahui kecepatan pemrosesan. Setial url diulangi sebanyak 10 kali. Setelah dijalankan maka didapatkan hasil berikut:

Tabel 2 Hasil Perhitungan 107225 Data URL

Proporsi Data	Hasil Akurasi (Rata - Rata)	
	Naïve Bayes	C4.5
65% - 35%	78.44%	86.79%
70% - 30%	78.46%	86.92%
75% - 25%	78.48%	86.87%
80% - 20%	78.41%	87.11%

Tabel 3 Hasil Perhitungan 100 Data URL

Proporsi Data	Hasil Akurasi (Rata - Rata)	
	Naïve Bayes	C4.5
65% - 35%	74.83%	98.00%
70% - 30%	72.60%	98.75%
75% - 25%	77.76%	98.25%
80% - 20%	74.17%	98.25%

Dari hasil percobaan diatas, terlihat bahwa hasil akurasi dari algoritma C4.5 lebih baik dibandingkan algoritma Naïve Bayes dengan perbedaan akurasi sekitar 9%. Dalam perhitungan akurasi terdapat 3 fitur yang mempengaruhi perhitungan, yaitu fitur ke 15, 18 dan 6.

Tabel 4 Waktu Pemrosesan Alamat URL

Url ke -	Rata - Rata Waktu Pemrosesan	
	Naïve Bayes	C4.5
1	14.18	13.23
2	21.29	18.07
3	61.77	58
4	18.93	18.26
5	27.51	25.53
6	16.91	14.98
7	18.87	17.56
8	18.99	17.55
9	15.94	14.53
10	18.72	20.1
Rata - Rata	23.31	21.78

Meskipun hasil yang didapatkan lebih tinggi, namun algoritma C4.5 memiliki waktu pemrosesan yang lebih cepat dibandingkan algoritma Naïve Bayes dengan perbedaan waktu pemrosesan sekitar dua detik. Fitur – fitur seperti fitur 6, 7 dan 13 yang keasliannya harus diambil dari server luar yaitu WHOIS.

4. Kesimpulan dan Saran

Kesimpulan yang didapat dari hasil pengujian yang telah dilakukan yaitu: Algoritma C4.5 dapat menghasilkan akurasi lebih baik dari algoritma lainnya, yaitu algoritma Naïve Bayes. Dengan tingkat akurasi pada data pengujian untuk data sebanyak 100 diraih sebesar 98.75%, data sebanyak 200 diraih sebesar 98.75%, data sebanyak 300 diraih sebesar 98.5% dan data sebanyak 107225 diraih sebesar 87.11%. Fitur pada website yang paling banyak mempengaruhi proses perhitungan adalah fitur keenam yaitu alamat url menggunakan “https” dan kejelasan alamat url terpecaya. Lama pemrosesan pencarian suatu alamat url baru sangat bervariatif, biasanya terkendala pada pencarian fitur – fitur seperti fitur keenam, ketujuh dan ketigabelas.

Beberapa saran yang dapat diberikan untuk upaya pengembangan program aplikasi lebih lanjut adalah: Melakukan pengembangan dari algoritma yang telah digunakan agar dapat mendapatkan nilai yang lebih baik. Pembuatan website ini diharapkan dapat menjadi suatu terobosan baru untuk pendeteksian url jahat dengan tipe lainnya, seperti malware, trojan dan spam.

REFERENSI

[1] Darma, Jarot S., dan Shenia A., 2009, “Buku Pintar Menguasai Internet”, Mediakita, Jakarta.

[2] Vyctoria., 2013, “Bongkar Rahasia E-Banking Security dengan Teknik Hacking dan Carding”, CV Andi Offset, Yogyakarta.

[3] Mohammad. Rami M., Thabtah. Fadi, and McCluskey. Lee., 2015, “Phishing Website Features”, School of Computing and Engineering, Huddersfield.

[4] IGN Mantra., 2015, “Potensi Ancaman Keamanan Email Perusahaan”, Info Komputer, Jakarta.

[5] S’to., 2011, “Certified Ethical Hacker 400% Illegal”, Jasakom, Jakarta.

[6] Efraim Turban, dkk., 2005, “Decision Support Systems and Intelligent Systems”, ANDI, Yogyakarta.

[7] Fayyad et al., 1996, “From data mining to knowledge discovery in databases”, AI Magazine, Vol.XVII, Nomor 3, h.16.

[8] Han, Jiawei., Kamber, Micheline dan Pei, Jian., 2012, “Data Mining Concepts and Techniques Third Edition”, Elsevier Inc, Amsterdam.

[9] Han, Jiawei, op.cit., h. 332

[10] Eko Prasetyo., 2012, “Data Mining: Konsep dan Aplikasi Menggunakan Matlab”, CV Andi Offset, Yogyakarta.

[11] PhishTank, https://www.phishtank.com/developer_info.php, 3 November 2019

[12] Phishstrom, https://research.aalto.fi/en/datasets/phishstorm--phishing--legitimate-url-dataset(f49465b2-c68a-4182-9171-075f0ed797d5).html, 9 November 2019,

[13] University of New Brunswick, https://www.unb.ca/cic/datasets/url-2016.html, 9 Novermber 2019

Kevin Marcello Jonathan, merupakan mahasiswa program Sarjana S1, program studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara.

Bagus Mulyawan, memperoleh gelar S.Kom dari Universitas Gunadarma pada tahun 1992. Kemudian memperoleh gelar M.M dari Universitas Budi Luhur pada tahun 2008. Saat ini aktif sebagai dosen tetap Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Tarumanagara, Jakarta.

Novario Jaya Perdana, memperoleh gelar S.Kom dari Institut Teknologi Sepuluh Nopember pada tahun 2011. Kemudian memperoleh gelar M.T. dari Universitas Indonesia pada tahun 2016. Saat ini aktif sebagai dosen tetap Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Tarumanagara, Jakarta.