Validasi Paket Data dengan Menggunakan Entropy dan Information Gain

Ahmad Fali Oklilas, Tasmi, Sri Desy Siswanti, Mira Afrina
Fakultas Ilmu Komputer
Universitas Sriwijaya
Palembang, Indonesia
fali@ilkom.unsri.ac.id, tasmi@ilkom.unsri.ac.id, siswanti_ctk@yahoo.com, mafrina@yahoo.com

***Abstrak*— Saat ini, internet lebih banyak menggunakan komunikasi terenkripsi dibandingkan kemananan (*security*). Saat ini sudah juga digunakan https dalam internet, tetapi ada permasalahan dalam hal klasifikasi paket terenskripsi. Penelitian ini merupakan pengembangan dari penelitian sebelumnya, dimana yang telah berhasil mengklasifikasi paket data secara *offline* tetapi belum berhasil dalam mengindentifikasi dan mengelompokkan layanan terinkripsi secara mendalam. Penelitian ini membangun simulasi secara nyata dalam pembuatan topologi dan klasifikasi dengan menggunakan *entropy* sebagai pemilihan atribut. Tujuan penelitian ini dapat melakukan validasi data untuk trafik terenkripsi dan *information gain*. Studi kasus dilakukan pada implementasi jaringan komputer di Fakultas Ilmu Komputer Unsri. *Feature* dengan metode *feature rangking* menghasilkan nilai *entropy* tertinggi 0.07414856 rangking pertama untuk atribut paket terenkripsi.**

Kata Kunci*— *paket data, keamanan jaringan, entropy, atribut

I. LATAR BELAKANG

Pada penelitian yang dilakukan oleh [1] menjelaskan trend pengunaan aplikasi di internet sudah menuju pada komunikasi terenkripsi dikarenakan memiliki kelebihan di bidang keamanan (*secure*), seperti aplikasi *sending tweets*, *messages to mobile*, *posting*, dan *search.* Penelitian oleh [2] menyatakan bahwa aplikasi HTTPS sudah banyak digunakan dalam internet, tetapi mereka juga menyatakan adalah permasalahan dalam hal klasifikasi paket terenskripsi.

Salah satu metode yang digunakan dalam mengenali paket yang lewat dalam sebuah network adalah *Deep Packets Inspection*, metode ini dapat melakukan inspeksi secara mendalam pada paket *header* TCP/IP dan *payload*. Penelitian sebelumnya yang dilakukan oleh [3] menggunakan metode DPI dalam mengelompokkan jenis trafik, sedangkan penelitian yang dilakukan oleh [4] membandingkan *tool* DPI dalam mengelompokkan jenis *traffic* internet.

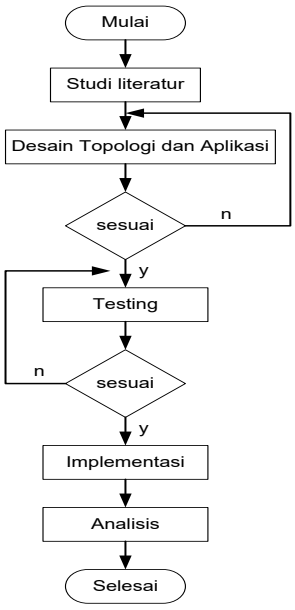
Solusi dalam mengklasfikasi trafik pada *network* telah banyak dilakukan dengan menghasilkan solusi yang aktif dan pasif sebagai solusi yang di tawarkan, seperti penelitian yang dilakukan oleh [5] menyatakan bahwa *management network* sebagai media pendukung dalam kasus identifikasi paket data dengan pendekatan *Operation, Administration, Maintenance & Provisioning* dan juga penelitian yang dilakukan oleh [6] mereka berhasil mengenali pola-pola paket dengan baik, namun sistem masih bersifat pasif sehingga tidak ada kontrol trafik yang keluar masuk.

Penelitian ini merupakan pengembangan dari penelitian sebelumnya oleh [3], dimana penelitian sebelumnya telah berhasil mengklasifikasi paket data secara *offline* tetapi belum berhasil dalam mengindentifikasi dan mengelompokkan layanan terinkripsi secara mendalam. Penelitian ini membangun simulasi secara nyata dalam pembuatan topologi dan klasifikasi sistem dengan menggunakan *entropy* sebagai pemilihan atribut.

Dengan semakin beragamnya aplikasi-aplikasi yang menggunakan komunikasi terenkripsi untuk pengamanan sebuah layanan, akan memunculkan permasalahan secara teknis yaitu: bagaimana mengenali dan mengelompokan jenis trafik terenkripsi. Pada penelitian ini diharapkan akan menghasilkan suatu sistem identifikasi paket dan *monitoring*. Fokus pada penelitian ini adalah bagaimana menghasilkan sebuah dataset yang digunakan untuk *feature extraction*, validasi data dan analisa data.

II. METODE PENELITIAN

Tahap-tahap dalam penelitian ini akan disajikan pada gambar 1 menggunakan diagram alir, dimana ada lima tahap yang diselesaikan dalam penelitian ini.

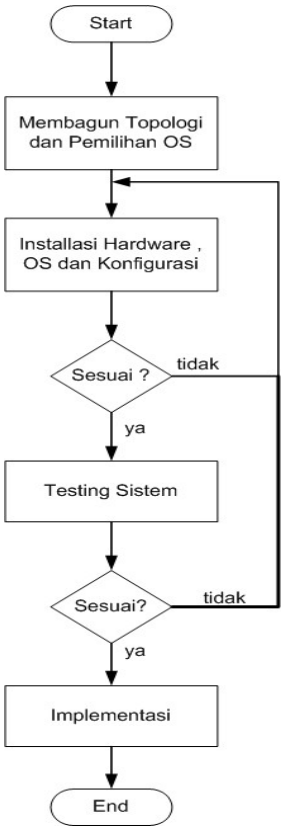


Gambar 1. Kerangka Kerja Penelitian

A. Network Design

Network design dilakukan dengan membuat penggambaran topologi, konfigurasi *router*, konfigurasi *switch*, alokasi IP Address, dan pemilihan sistem *monitoring* dan identifikasi. Hasil dari rancangan akan digunakan untuk

implementasi di tahapan selanjutnya. Gambar 2 merupakan diagram alir *Network Design*.



Gambar 2. Diagram Alir *Network Design*

B. *Testing Konfigurasi dan Deamon yang digunakan*

Tahap berikutnya adalah testing. Adapun percobaan yang dilakukan adalah Ping sistem baik dari dari dalam maupun dari luar, *browsing* ke *public* (internet), *remote control* dari jarak jauh menggunakan SSH. Jika hasil tidak sesuai yang diharapkan maka akan mengerjakan kembali langkah kedua.

C. *Implementasi*

Pada tahapan ini akan dibuat *prototype*, dalam kasus ini akan dilakukan implementasi di jaringan komputer Fakultas Ilmu Komputer Unsri. Adapun detail pekerjaan pada tahapan ini adalah: topologi untuk tahap awal dalam *prototype*, topologi terintegrasi ke dalam jaringan nirkable, *coding* aplikasi dan konfigurasi, integrasi perangkat kedalam aplikasi, analisis.

Topologi jaringan klasifikasi trafik terenkripsi yang dibangun dapat dijadikan sebagai dataset untuk mengembangkan aplikasi pengklasifikasian paket data, dimana jaringan tersebut sudah dapat mewakili secara umum perangkat yang digunakan dalam sebuah pengklasifikasian paket data. (1) satu buah *router* sebagai *forwarder* paket data di *layer* 3 *network*, (2) satu buah *switch layer* 2 *datalink*, (3) tiga buah PC *Workstation user* yang melakukan *behaviour activity*.

D. *Pengumpulan Data*

Pada penelitian ini paket data diambil secara *real-time* dengan membangun jaringan komputer di kampus Fasilkom Unsri dengan dengan menggunakan IP Publik. Data yang akan yang akan ditangkap dalam penelitian ini adalah yang normal dan paket data terenkripsi dalam jaringan yang dibangun.

E. *Feature Extraction*

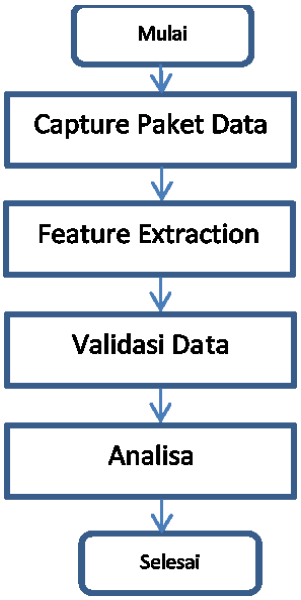
Pada tahap ini adalah proses bagaimana dapat mengambil informasi-informasi atau atribut-atribut dari *data capture traffic* yang dibutuhkan. Pengambilan atau proses mengekstrak file *pcap* atau *tcpdump* mengunakan program.

Ada dua proses yang dilakukan di bagian ini yaitu (i) *capture Packet*, (2) *extract Packet data*. *Capture packet* sangat dibutuhkan dalam penelitian ini untuk melihat paket-paket yang lewat dalam *network* yang dibangun. *Extract Packet* ini adalah proses untuk mendapatkan informasi-informasi dan nilai-nilai atribut dari *data capture traffic*.

Tujuan dari *feature extraction* adalah memetakan pola berdasarkan ciri-ciri yang dimiliki oleh suatu *packet*, maka klasifikasi bertujuan untuk mengenali *packet* dengan cara mengklasifikasikan ciri-ciri yang dimilikinya. Pada penelitian ini *feature extraction* akan menggunakan *wincap* dan *phyton* sabagai bahasa pemograman.

F. *Model Klasifikasi Trafik Terenkripsi*

Pada bagaian ini diusulkan metode untuk klasifikasi trafik data terenkripsi dengan menggunakan algoritma *patterns matching* dengan pendekatan *automaton-based* secara *real-time*. Ada tiga tahap yang akan dilakukan pada kegiatan ini seperti pada gambar 3. Proses pertama dalam *preprocessing* yang terdiri dari proses *capture packets*, *featureselection*.

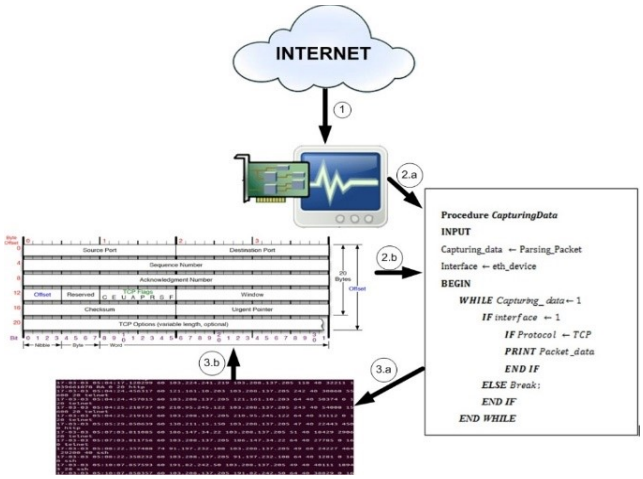


Gambar 3. Kerangka Kerja Klasifikasi

III. HASIL DAN PEMBAHASAN

A. *Capture Packet Data*

Sniffing paket adalah sebuah aplikasi yang digunakan untuk memantau trafik pada sebuah jaringan melintas. Penelitian ini diawali mendesain sebuah topologi yang digunakan untuk menghasilkan dataset dengan menggunakan library libpcap seperti pada gambar 4.



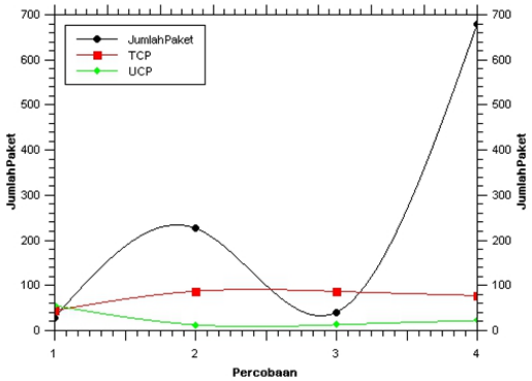
Gambar 4. Proses *Capture Paket Data*

Dari gambar 4, untuk menghasilkan sebuah *dataset* secara *real*, maka dibuat sebuah algoritma dengan menggunakan *library pcap*. Tahap 1 membangun sebuah hardware untuk media *capture* paket data, kemudian pada tahap 2 adalah membangun sebuah algortima untuk membaca *raw data* yang ditangkap oleh NIC. Pada proses ini (2a) hal pertama yang dilakukan adalah mendefenisikan *interface* yang digunakan, *protocol* yang akan ditangkap serta jumlah paket yang ditangkap setiap detik. Poin 2b adalah gambar *header* dari tcp/ip yang akan ditangkap. Proses 3a merupakan hasil dari *capture* secara *real* yang menghasilkan *dataset* yang akan digunakan pada tahap selanjutnya. Tabel 1 merupakan hasil dari *capture* paket data.

Tabel 1. Hasil *Capture* Paket Data

No	Percobaan	Jumlah Paket	TCP	UDP	Paket Drop
1	Ke -1	29.164	44.59%	55.26%	0.15%
2	Ke -2	226.516	87.11%	12.78%	0.11%
3	Ke -3	40.659	86.82%	13.05%	0.12%
4	Ke -4	679.320	77.53%	22.41%	0.05%

Pada proses pengambilan data sebanyak empat kali didapatkan bahwa data pada percobaan ke empat menghasil data yang baik dikarena paket yang hilang hanya sekitar 0.05%, paket TCP didapat 77.53%, dan Paket UDP 22.41%. pada percobaan pertama didapatkan paket UDP lebih besar dibandingkan data TCP sebesar 55.26% perbanding 44.59%, hal ini dikarena pada percobaan pertama beberapa protocol UDP seperti DHCP, DNS dan Netbios masih melakukan *broadcast*.



Gambar 5. Hasil Capture TCP dan UDP

Gambar 5 menampilkan hasil *capture* berdasarkan aplikasi-aplikasi yang diakses oleh user. Pada algortima yang dibangun hanya memfokuskan pada protokol TCP dan UDP. Sedangkan beberapa paket seperti ICMP, SIP, SNMP dan lain-lain tidak diproses.

B. Hasil Feature Extraction

Bagian ini akan membaca atribut-atribut yang dihasil dari hasil *capture*, dimana hasil *capture* menghasilkan *raw data* yang sulit dibaca oleh manusia ini dikarenakan *header* IPV4 memiliki susunan yang unik dan juga adanya proses *encapsulated* paket data. Oleh karena itu membaca data dari *raw data* dibangunlan sebuah algoritma yang berfungsi untuk mengekstraksi paket data tersebut, dengan tujuan untuk mendapat nilai-nilai dari semua atribut serta hasil *features extraction* diubah ke dalam bentuk file.csv yang akan digunakan untuk proses *training* karena dengan tipe file ini lebih mudah dalam proses *training*.

Hasil validasi antara program *network monitoring* dengan algoritma yang dibangun memiliki beberapa atribut berhasil diekstraksi. Hasil algoritma ektraksi menghasikan delapam belas atribut yang terdiri dari : (1) Source_IP, (2) DST_IP, (3) TTL, (4) Capture_length, (5) Header_Length, (6) Total_length, (7) Idetification, (8) Frag_Offset, (9) Source_Port, (10) DST_Port, (11) Flags, (12) Windows, (13) Urgent_Pointer, (14) Ack, (15) Checksum_Header, (16) protocol, (17) Seq dan (18) Checksum_Protokol.

Hasil yang didapat dari hasil FE menampilkan bahwa *user* dengan alamat 10.100.115.7 sedang melakukan proses mengakses aplikasi google.com dengan Alamat 172.217.27.3. Aplikasi google.com merupakan salah satu layanan yang menggunakan data terenkripsi dibuktikan dari hasil yang didapat yaitu Source Port nya adalah 443 dan juga menggunakan *protocol* TLS (*Transport Layer Security*) ver 1.2, dimana *protocol* kriptografi untuk menyediakan komunikasi yang aman melalui internet.

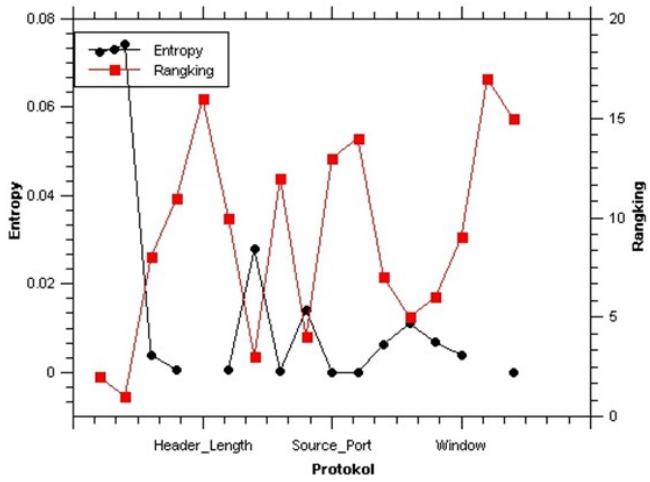
C. Hasil Feature Selection

Pada bagian ini ada dua pekerjaan utama yang dilakukan. Pertama adalah menentukan atribut-atribut hasil dari trafik yang ditangkap secara *real-time*, dan yang kedua adalah menentukan atribut-atribut yang relevan dari hasil *capture* yang didapat dari trafik di Fasilkom Unsri dengan menggunakan metode feature rangking yaitu *Information Gain* (IG) untuk mencari rangking yang akan divalidasi dengan menggunakan metode klasifikasi *naive bayes*. Hasil dari klasifikasi akan dianalisa berdasarkan tingkat akurasi dari setiap atribut.

Bagian pertama yang dilakukan adalah *pre-processing* data yang didapat. Pada tahap ini adalah proses pembersihan data yang digunakan untuk menghilangkan data-data yang *error* sehingga didapat hasil akurasi yang baik. Tahap selanjutnya adalah membagi *dataset* menjadi dua bagian yaitu *training* sebesar 70% dan *testing* sebesar 30%. Hal ini digunakan untuk proses pembelajaran untuk mendapat prediksi atribut-atribut yang kuat dalam menentukan pola sebuah paket data maupun sebuah serangan. Tabel 2 dan gambar 6 menampilkan hasil *entropy* dan rangking dari setiap atribut dengan menggunakan metode information Gain (IG).

Tabel 2. Hasil Rangking Atribut dengan menggunakan IG

No	Atribut	IG	
		Entropy	**Rangking**
1	Protokol	0,0000	18
2	Source_IP	0.07236785	2
3	DST_IP	0.07414856	1
4	TTL	0.00397622	8
5	Capture_length	0.00055618	11
6	Header_Length	0,0000	16
7	Total_Lenght	0.00055618	10
8	Identification	0.02781127	3
9	Checksum_Header	0.00025073	12
10	Frag_Offset	0.01413719	4
11	Source_Port	0.00002238	13
12	DST_Port	0.00002238	14
13	Flags	0.0063704	7
14	Ack	0.01102643	5
15	Seq	0.00676296	6
16	Window	0.0039587	9
17	Urgent_Pointer	0,0000	17
18	Checksum_Protokol	0.00000774	15



Gambar 6. Hasil Rangking dan Nilai *Entropy* Atribut

Hasil rangking didapatkan bahwa atribut Source_port merupakan rangking pertama ini artinya atribut ini mempunyai kontribusi yang besar terhadap atribut *service* yang dijadikan class, sedangkan tiga atribut yaitu Header_length, Urgent_Pointer dan Protokol merupakan tiga rangking terendah, ini artinya atribut-atribut ini tidak mempunyai kontribusi.

IV. KESIMPULAN

Penelitian ini bertujuan untuk mengklasifikasi layanan terenskripsi. Penelitian ini membangun simulasi secara nyata dalam pembuatan topologi dan klasifikasi dengan menggunakan *entropy* sebagai pemilihan atribut. Tujuan penelitian ini dapat melakukan validasi data untuk trafik terenkripsi dan *information gain*. Studi kasus dilakukan pada implementasi jaringan komputer di Fakultas Ilmu Komputer Unsri. Pada proses *capture packet data*, *packet loss* dapat dikurangi dengan cara memperbesar *range buffer*. Hasil *feature selection* dengan menggunakan metode *feature rangking* didapatkan nilai *entropy* tertinggi adalah 0.07414856 untuk atribute DST_IP.

DAFTAR PUSTAKA

[1] L. Deri, M. Martinelli, T. Bujlow, and A. Cardigliano, "NDPI: Open-source high-speed deep packet inspection," in *IWCMC 2014 - 10th International Wireless Communications and Mobile Computing Conference*, 2014, pp. 617–622.

[2] M. Husák, M. Cermák, T. Jirsík, and P. Celeda, "Open Access HTTPS traffic analysis and client identification using passive SSL / TLS fingerprinting," *EURASIP J. Inf. Secur.*, 2016.

[3] A. F. Oklilas and Tasmi, "Monitoring and Indentification Packet in Wireless With Deep Packet Inspection Method," *Int. Conf. Recent Trends Phys. 2016 IAES Int. Conf. Electr. Eng. Comput. Sci. Informatics*, vol. 365, p. 011001, 2017.

[4] T. Bujlow, V. Carela-Español, and P. Barlet-Ros, "Independent comparison of popular DPI tools for traffic classification," *Comput. Networks*, vol. 76, pp. 75–89, 2015.

[5] M. Molina, I. Paredes-oliva, W. Routly, and P. Barlet-ros, "Operational experiences with anomaly detection in backbone networks," *Comput. Secur.*, vol. 31, no. 3, pp. 273–285, 2012.

[6] H. Zhang, D. Yao, N. Ramakrishnan, and Z. Zhang, "Causality reasoning about network events for detecting stealthy malware activities," *Comput. Secur.*, vol. 58, pp. 180–198, 2016.