

**DESAIN DAN IMPLEMENTASI SISTEM *CYBER THREAT  
INTELLIGENCE* BERBASIS *WEB DATA EXTRACTION* PADA  
REDDIT**

**Laporan Tugas Akhir**

**Disusun sebagai syarat kelulusan tingkat sarjana**

**Oleh**

**FATHAN ANANTA NUR**

**NIM : 18219008**



**PROGRAM STUDI SISTEM DAN TEKNOLOGI INFORMASI  
SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA  
INSTITUT TEKNOLOGI BANDUNG  
JULI 2023**

**DESAIN DAN IMPLEMENTASI SISTEM *CYBER THREAT INTELLIGENCE* BERBASIS *WEB DATA EXTRACTION* PADA  
REDDIT**

**Laporan Tugas Akhir**

**Oleh**

**FATHAN ANANTA NUR**

**NIM : 18219008**

**Program Studi Sistem dan Teknologi Informasi**

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

Telah disetujui dan disahkan sebagai Laporan Tugas Akhir  
di Bandung, pada tanggal 24 Februari 2023

Pembimbing,

Ir. Budi Rahardjo, M.Sc., Ph.D.

NIP 109110001

## **LEMBAR PERNYATAAN**

Dengan ini saya menyatakan bahwa:

1. Pengerjaan dan penulisan Laporan Tugas Akhir ini dilakukan tanpa menggunakan bantuan yang tidak dibenarkan.
2. Segala bentuk kutipan dan acuan terhadap tulisan orang lain yang digunakan di dalam penyusunan laporan tugas akhir ini telah dituliskan dengan baik dan benar.
3. Laporan Tugas Akhir ini belum pernah diajukan pada program pendidikan di perguruan tinggi mana pun.

Jika terbukti melanggar hal-hal di atas, saya bersedia dikenakan sanksi sesuai dengan Peraturan Akademik dan Kemahasiswaan Institut Teknologi Bandung bagian Penegakan Norma Akademik dan Kemahasiswaan khususnya Pasal 2.1 dan Pasal 2.2.

Bandung, 24 Februari 2023

Fathan Ananta Nur

NIM 18219008

## **ABSTRAK**

# **DESAIN DAN IMPLEMENTASI SISTEM CYBER THREAT INTELLIGENCE BERBASIS WEB DATA EXTRACTION PADA REDDIT**

Oleh

FATHAN ANANTA NUR

NIM : 18219008

Ancaman siber telah merambat menjadi masalah sosial. Perusahaan telah banyak berinvestasi dalam pertahanan siber dalam upaya memerangi hal ini. *Cyber Threat Intelligence* (CTI) adalah salah satu mekanisme pertahanan yang akan dibahas. Organisasi baru-baru ini mulai melakukan investasi yang signifikan dalam pengembangan CTI untuk memerangi meningkatnya ancaman serangan siber. CTI dicirikan sebagai sekelompok data faktual yang mencakup konteks, metode ancaman, indikator, dan penanggulangan potensial. Komunitas dan forum *hacker* atau *bad actor* dapat memberikan CTI proaktif yang substansial. Forum, jika dibandingkan dengan platform lain, menawarkan metadata terlengkap, data permanen, dan puluhan *tools*, *technique*, dan *procedure* (TTP) yang dapat diakses secara terbuka. Salah satu platform yang mewadahi komunitas dan forum adalah Reddit. Reddit menyediakan izin untuk mengambil data-data esensial yang dapat dikonversikan menjadi CTI. Praktisi profesional dapat lebih mudah memutuskan tindakan masa depan mereka dengan menggunakan kesimpulan analisis CTI sebagai bantuan visual. Oleh karena itu, data-data yang dikumpulkan dari Reddit dapat diolah dan diberikan visualisasi sesuai konteks, sehingga dapat membantu perusahaan maupun organisasi menyiapkan langkah preventif terkait ancaman siber.

Kata kunci: CTI, Reddit, *web scraping*, *data visualization*

## KATA PENGANTAR

Puji syukur kehadiran Tuhan Yang Maha Esa, sehingga penulis dapat menyelesaikan laporan tugas akhir ini. Laporan dengan judul “Desain dan Implementasi Sistem *Cyber Threat Intelligence* Berbasis *Web Data Extraction* pada Reddit” disusun guna memenuhi persyaratan menyelesaikan mata kuliah Tugas Akhir dan persyaratan kelulusan di Program Studi Sistem dan Teknologi Informasi, Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung. Dalam penyusunan laporan tugas akhir, tentunya penulis mendapatkan pengetahuan dan pengalaman dari beberapa pihak. Oleh karena itu, penulis mengucapkan terima kasih kepada:

1. Yudistira Dwi Wardhana Asnar, S.T., Ph.D., selaku Kepala Program Studi Sistem dan Teknologi Informasi,
2. Dr. Fetty Fitriyanti Lubis, S.T., M.T., selaku koordinator Mata Kuliah Tugas Akhir Program Studi Sistem dan Teknologi Informasi,
3. Ir. Budi Rahardjo, M.Sc., Ph.D., selaku dosen pembimbing tugas akhir yang selalu memberikan arahan dan inspirasi,
4. Muhammad Aris Kusnanta dan Nur Hayati selaku orang tua penulis yang selalu memberikan dukungan paling utama dalam penyelesaian tugas akhir,
5. dan segenap pihak yang terlibat dalam pelaksanaan dan penyusunan laporan tugas akhir, sehingga laporan tugas akhir dapat terselesaikan dengan baik.

Penulis menyadari bahwa laporan tugas akhir ini masih jauh dari kata sempurna. Oleh karena itu, kritik dan saran selalu penulis harapkan, demi penyusunan laporan yang lebih baik lagi kedepannya. Penulis berharap, semoga laporan tugas akhir ini dapat bermanfaat untuk penulis sendiri, dan para pembaca.

## DAFTAR ISI

<b>BAB I PENDAHULUAN.....</b>	<b>1</b>
I.1    Latar Belakang.....	1
I.2    Rumusan Masalah.....	3
I.3    Tujuan .....	4
I.4    Batasan Masalah .....	4
I.5    Metodologi.....	5
I.6    Sistematika Pembahasan.....	6
<b>BAB II STUDI LITERATUR .....</b>	<b>7</b>
II.1    Cyber Threat Background.....	7
II.2    CTI Sharing.....	9
II.3    Model CTI.....	10
II.3.1    Post-Event CTI Sharing .....	11
II.3.2    Pre-Event CTI Sharing .....	11
II.4    Ekstraksi Informasi .....	12
II.5    Forum Internet .....	13
II.6    Reddit.....	14
II.7    Penelitian Terkait.....	17
II.7.1    Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics [PRO21].....	17
II.7.2    Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence [SAM17].....	17
<b>BAB III DESAIN DAN IMPLEMENTASI .....</b>	<b>19</b>

III.1	Analisis Permasalahan.....	19
III.2	Rancangan Solusi Secara Garis Besar .....	23
III.3	Linimasa Penyelesaian Tugas Akhir .....	26
<b>BAB IV IMPLEMENTASI DAN EVALUASI.....</b>		<b>28</b>
IV.1	Ekstraksi Data.....	28
IV.1.1	Reddit Developer API.....	28
IV.1.2	Python Reddit API Wrapper (PRAW).....	29
IV.1.3	Struktur Data.....	30
IV.2	Virtual Machine.....	35
IV.2.1	Google Cloud Platform (GCP) .....	35
IV.2.2	NoMachine Remote Desktop.....	37
IV.3	Pengujian Kemampuan Crawling Sistem.....	37
IV.3.1	Berdasarkan Network .....	37
IV.3.2	Berdasarkan Paket yang Diminta.....	39
IV.4	Persiapan Data .....	42
IV.4.1	Pembersihan Data .....	42
IV.4.2	Manajemen Stop Word .....	43
IV.5	Visualisasi Data .....	44
<b>BAB V KESIMPULAN DAN SARAN .....</b>		<b>45</b>
V.1	Kesimpulan .....	45
V.2	Saran .....	45

## **DAFTAR LAMPIRAN**

<b>Lampiran A. Contoh Judul Lampiran.....</b>	<b>50</b>
A.1 Contoh Judul Anak Lampiran.....	50



## DAFTAR GAMBAR

Gambar II.1. Tahapan konstruksi koleksi retorik kalimat **Error! Bookmark not defined.**

## DAFTAR TABEL

Tabel II.1. Pengelompokan *Tag* MARC-21 .....**Error! Bookmark not defined.**

# **BAB I**

## **PENDAHULUAN**

Bab Pendahuluan pada laporan ini dijadikan sebagai landasan kerja dan arah kerja tugas akhir yang berfungsi untuk mengantarkan pembaca dalam membaca laporan tugas akhir secara keseluruhan.

### **I.1 Latar Belakang**

Kejahatan siber telah berkembang secara signifikan sejak komputer dikembangkan dan dibangun untuk dapat berkomunikasi satu sama lain. Bertepatan dengan paradigma Revolusi Industri 4.0, semakin banyak perusahaan yang menghubungkan pabrik dan infrastruktur bisnis mereka ke internet, yang juga disebut *Industrial Internet*, untuk meningkatkan efektivitas dan efisiensinya. Didorong oleh kekhawatiran yang berkembang akan potensi *vulnerability* pada jaringan dan oleh meningkatnya jumlah gangguan di domain siber, banyak organisasi dan perusahaan mengambil langkah-langkah untuk lebih memahami *vulnerability* dan ancaman yang menjadi sasaran infrastruktur informasi mereka. Terlebih dari itu, banyak organisasi telah memutuskan mengambil langkah-langkah untuk melindungi aset-aset tersebut. Untuk melawan meningkatnya ancaman serangan siber, organisasi dalam beberapa tahun terakhir telah mulai fokus berinvestasi dalam mengembangkan *Cyber Threat Intelligence* (CTI).

CTI pada umumnya merupakan proses yang berbasis data dengan mengumpulkan dan menganalisis data dari sistem internal seperti informasi keamanan dan sistem manajemen *event*, *file log*, sistem deteksi dan pencegahan intrusi jaringan, dan lainnya untuk memberikan wawasan tentang ancaman yang muncul dan aktor ancaman siber. Terlepas dari nilai dan prevalensinya, data yang dikumpulkan dari sistem internal dianggap CTI reaktif. Untuk mendapatkan data-data sistem internal seperti pada CTI reaktif, diperlukan peristiwa serangan maupun gangguan sebagai objek yang ditinjau. Hal ini tentunya memerlukan pengorbanan

tersendiri dari sistem. Untuk mengatasi hal tersebut, perusahaan memerlukan sistem CTI tanpa adanya serangan ataupun gangguan yang masuk ke dalam sistem. Oleh karena itu, istilah CTI proaktif dikenalkan sebagai penyanggah CTI reaktif.

Dibandingkan dengan CTI reaktif, CTI proaktif cenderung bersifat preventif dan dilakukan sebelum terjadinya gangguan serangan yang menyebabkan kerugian pada suatu sistem. Informasi mengenai CTI dapat ditemukan dimana saja. Untuk CTI proaktif, informasi secara aktif akan dicari di entitas luar. Terdapat banyak entitas luar yang disinyalir mengandung banyak informasi yang dinilai dapat membahayakan organisasi dan sistem. Media sosial belakangan ini menjadi tren yang digunakan masyarakat umum. Aktor jahat yang secara anonim tergabung ke dalam struktur masyarakat dapat mengakses ke media sosial dan berinteraksi dengan pihak lainnya.

Salah satu penggunaan media sosial yang cukup banyak dilakukan saat ini adalah forum diskusi online. Forum diskusi online berbentuk papan pesan online yang dibuat agar para anggota forum dapat berkomunikasi satu sama lain dengan mendiskusikan berbagai topik sehingga memungkinkan untuk saling bertukar pikiran dan pengetahuan [WIL18]. Informasi esensial dan sensitif yang menjadi ancaman bagi organisasi terkadang juga dapat dipertukarkan dan mengalir di dalam diskusi online. Sama seperti media sosial lainnya, forum diskusi dapat diakses melalui platform yang bervariasi, salah satu platform yang mudah diakses adalah *website*. Oleh karena itu, sistem *web data scraper* dapat diimplementasikan untuk melintasi struktur tubuh web dan mengumpulkan isinya. Perangkat lunak yang dikenal sebagai *web crawler* menjelajahi internet dengan mengklik tautan dan mengumpulkan halaman web menggunakan protokol HTTP [FU10].

Salah satu forum diskusi online adalah Reddit. Reddit adalah media sosial berjenis agregasi berita, pemeringkatan konten, dan situs web diskusi. Pengguna terdaftar dapat mengirimkan konten ke Reddit berupa tautan, *posting* teks, gambar, dan video, yang kemudian dilihat dan dipilih oleh anggota lain. Postingan diatur berdasarkan subjek ke dalam papan diskusi yang dibuat pengguna yang disebut komunitas atau *subreddits*. Di Indonesia sendiri, Reddit merupakan situs yang

diblokir oleh pemerintah dikarenakan konten yang ada di dalamnya dinilai terlalu bebas dan sangat bervariasi termasuk konten pornografi yang dilarang di Indonesia. Dilansir [similarweb.com](https://www.similarweb.com), pada November 2022, Reddit menempati posisi ke-20 *website* yang paling sering dikunjungi di seluruh dunia. Oleh karena itu, Reddit dinilai memiliki banyak informasi yang lengkap dan sangat bervariasi baik kontennya maupun penggunanya. Model CTI proaktif cukup cocok jika diimplementasikan ke Reddit karena akan memberikan banyak insight mengenai model dan aktor ancaman ke suatu sistem.

## **I.2 Rumusan Masalah**

Pengembangan sistem intelijen dari data internal adalah fokus utama dari inisiatif CTI saat ini. Hal ini menunjukkan bahwa langkah-langkah keamanan saat ini sering dikelola secara reaktif daripada proaktif. Selain itu, sifat website suatu forum membuat metode *web scraper* untuk mengambil data kurang efektif untuk navigasi dan pengindeksan yang efisien. Sehingga, upaya yang dilakukan untuk mendapatkan data dari website forum diskusi, sebagian besar terkonsentrasi pada pengumpulan *batch* dan pemrosesan data forum peretas. Koleksi statis ini menjadi kurang bernilai saat ancaman berubah seiring waktu. Maka dengan itu, masalah yang diambil pada artikel ini meliputi

- a. Apakah model *web scraper* dapat diterapkan untuk mengambil dan mengolah data yang dapat dimungkinkan untuk menyusun CTI proaktif?
- b. Apa saja informasi dan *insight* berharga yang bisa diperoleh dari Reddit dengan *web scraping*?
- c. Bagaimana kinerja model *web scraper* pada Reddit dengan mempertimbangkan masalah yang dihadapi saat pengimplementasian untuk terus mengumpulkan informasi dan *insight* esensial secara kontinu?
- d. Bagaimana visualisasi interaktif dikembangkan dan disajikan kepada akademisi dan praktisi CTI untuk menyelidiki eksploitasi yang dikumpulkan untuk CTI proaktif dan tepat waktu?

### **I.3 Tujuan**

Penelitian ini dilakukan untuk membuktikan dan memberikan jawaban dari rumusan masalah yang sudah tertulis sebelumnya. Tujuan dari penulisan artikel ini adalah

- a. Membangun model *web scraper* yang dapat diterapkan untuk mengambil dan mengolah data yang dapat dimungkinkan untuk menyusun CTI proaktif.
- b. Mengetahui informasi dan *insight* berharga yang bisa diperoleh dari Reddit dengan *web scraping*.
- c. Mengetahui kinerja model *web scraper* pada Reddit dengan mempertimbangkan masalah yang dihadapi saat pengimplementasian untuk terus mengumpulkan informasi dan *insight* esensial secara kontinu.
- d. Membuat visualisasi interaktif disajikan kepada akademisi dan praktisi CTI untuk menyelidiki eksploitasi yang dikumpulkan untuk CTI proaktif dan tepat waktu.

### **I.4 Batasan Masalah**

Penelitian ini meliputi pembuatan suatu model *web scraper* yang difungsikan untuk mengambil data yang akan diolah menjadi *insight* CTI yang dapat dianalisis dan digunakan sebagai dasar pengambilan tindakan. Pengimplementasian model *web scraper* dilakukan pada forum diskusi online Reddit. Platform yang dipilih merupakan website Reddit sehingga *web scraper* dapat berjalan di atasnya. Penelitian ini menghasilkan suatu *prototype* yang dapat mensimulasikan cara kerja *web scraper* pada Reddit. Fokus topik yang akan dilakukan pengambilan data CTI dari Reddit meliputi kasus kebocoran data Tokopedia dan informasi atau diskusi umum mengenai sistem perbankan di Indonesia. Pembangunan model data scraper pada Reddit mengikuti metodologi dan fase-fase *Software Development Life Cycle* (SDLC) sehingga mudah untuk melakukan evaluasi pada setiap langkahnya.

## I.5 Metodologi

Metode yang digunakan dalam penyusunan sistem ini adalah *design thinking*. *Design thinking* merupakan salah satu metodologi desain yang memberikan pendekatan berdasarkan solusi untuk menyelesaikan masalah. Metode ini dipilih karena pendekatan *design thinking* sangat mengandalkan solusi untuk menjawab suatu permasalahan yang diangkat. Pendekatan semacam ini akan menuntut proses untuk memunculkan sesuatu yang konstruktif demi mengatasi sebuah masalah.

Design thinking meliputi 5 langkah yang berurutan. Tahap pertama dari proses *design thinking* adalah *empathize* yang memiliki fungsi untuk mendapatkan pemahaman dan empati tentang masalah yang akan diselesaikan. Langkah ini dilakukan untuk mengenali secara menyeluruh tentang permasalahan yang ada. Tahap kedua dalam melakukan *design thinking* adalah *define*. Dalam tahap ini, informasi yang diperoleh dari tahap sebelumnya diolah. Di sinilah informasi tersebut akan dianalisis dan disintesis untuk mengidentifikasi masalah utama yang akan ditemukan solusinya. Tahap ini menghasilkan pernyataan masalah secara final. Tahap ketiga dari proses *design thinking* adalah *ideate*. Proses *ideate* ini ditandai dengan bermunculannya ide-ide awal dari masalah yang sudah ditentukan sebelumnya. Pada akhir tahap ini, diputuskannya ide terbaik yang akan dibawa untuk memecahkan masalah supaya dapat mempermudah dalam penyusunan pengujian. Tahap selanjutnya adalah *prototype* yang menghasilkan serangkaian versi produk yang diperkecil dan dibatasi atau fitur spesifik yang ditemukan dalam solusi. Solusi diimplementasikan dalam *prototype* akan diteliti dan diperiksa sehingga menghasilkan suatu keputusan bahwa solusi tersebut akan diterima untuk masuk tahap selanjutnya ataupun direvisi dan diperiksa kembali. Tahap testing adalah tahap akhir dari design thinking, tetapi memungkinkan untuk keseluruhan proses design thinking dapat berulang, hasil yang dihasilkan dalam tahap pengujian sering digunakan untuk mendefinisikan kembali satu atau lebih masalah yang ditemukan di tengah proses.

## **I.6 Sistematika Pembahasan**

Laporan ini terdiri dari lima bab. Bab pertama adalah pendahuluan yang berisi subbab sebagai berikut: latar belakang, rumusan masalah tujuan, batasan masalah, metodologi, dan sistematika pembahasan. Bab kedua merupakan studi literatur yang secara garis besar berisi teori-teori terkait dan penelitian terkait. Bab ketiga merupakan desain dan implementasi CTI berbasis *web scraper* pada Reddit. Bab ketiga, secara garis besar, berisi analisis masalah, desain solusi, dan timeline penyelesaian tugas akhir. Bab keempat berisikan evaluasi dari solusi yang diajukan di bab ketiga. Bab kelima merupakan kesimpulan yang menjawab permasalahan utama pada laporan tugas akhir.



## **BAB II**

### **STUDI LITERATUR**

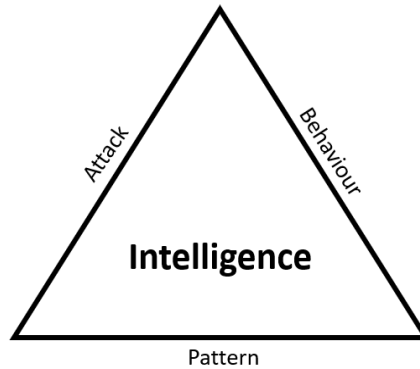
Bab ini mengemukakan landasan metode, teknik, maupun teori yang mendasari, terkait atau yang digunakan pada permasalahan yang dikaji dalam pembangunan Sistem CTI Berbasis *Web Scraper* pada Reddit. Selain itu, bab ini juga menyajikan penelitian serupa yang pernah dilakukan untuk menjaga ketersinambungan ilmu pengetahuan dan *state of the art*.

#### **II.1 Cyber Threat Background**

Seiring berjalannya waktu, entitas organisasi akan selalu menerima tantangan yang signifikan, sehingga diharuskan bagi mereka untuk mempertahankan data dan sistem mereka. Tantangan tercipta dari peningkatan frekuensi dan kecanggihan serangan siber yang diciptakan oleh *threat actor* yang cakap. *Threat actor* yang biasanya bersifat ambisius dan gesit dapat menggunakan berbagai taktik, teknik, dan prosedur (TTP) untuk membahayakan sistem, mengganggu layanan, melakukan tipuan, dan mengungkapkan atau mencuri kekayaan intelektual dan informasi sensitif lainnya [JOH16]. *Threat actor* dapat terdiri dari lingkup individu hingga kelompok sumber daya, bertindak secara sistematis sebagai bagian dari entitas kriminal sampai atas nama negara.

Serangan siber yang dibawa oleh aktor negara dapat menyebabkan kasus politik, diplomatik antar negara yang terlibat dan bahkan menyebabkan kerugian yang lebih besar. Sebagai contoh kasus, pada tanggal 23 Desember 2015, Ukrainian Kyivoblenergo, sebuah perusahaan distribusi listrik regional di Ukraina, melaporkan pemadaman layanan listrik kepada pelanggan. Pemadaman ini disebabkan oleh pihak ketiga yang masuk secara ilegal ke komputer perusahaan dan sistem mereka. Tak lama setelah serangan itu, pejabat pemerintah Ukraina mengklaim pemadaman itu disebabkan oleh serangan siber, dan bahwa dinas keamanan Rusia bertanggung jawab atas insiden tersebut [LEE16].

Berikut ini adalah segitiga konseptual pengenalan serangan siber. Ada tiga komponen utama: *attack*, *behaviour*, dan *pattern* [ALM18].



**Gambar 1 Segitiga Konseptual Serangan Siber**

Ide utama segitiga pada Gambar 1 adalah bahwa kumpulan data insiden siber berisi data serangan, yang dapat dianalisis menggunakan analisis data. Data serangan dapat dipisahkan dari insiden normal dan disajikan dalam format yang lebih mudah dibaca. Jadi, dalam konteks ini serangan yang dilakukan oleh penyerang mengungkapkan perilaku penyerang. Dengan menambahkan aspek intelijen seperti analisis data, pada dua komponen ini kita dapat mengidentifikasi pola serangan. Pola serangan bisa menjadi kunci untuk mencegah serangan siber di masa depan.

Umumnya, pengumpulan data log sistem dilakukan untuk sebagian besar sistem. Data tersebut dapat dikatakan sebagai *big data* karena data tersebut memenuhi konsep *velocity*, *verity* dan *volume* [HIL15]. Jika yang dipertimbangkan hanya volume kumpulan data, maka akan membutuhkan teknik khusus untuk menganalisis dan menyajikannya. Dengan menganalisis data ini, dapat diidentifikasi peristiwa serangan. Peristiwa serangan ini dapat terjadi berulang kali dari waktu ke waktu, yang dapat membentuk pola. Tujuan menggunakan analisis data adalah untuk mengidentifikasi pola seperti itu. Data dapat dianalisis lebih cerdas dan efisien dengan menggunakan teknik analisis *big data*.

## II.2 CTI Sharing

Adanya perkembangan digital yang cepat, bidang serangan yang meluas, dan semakin banyak kerentanan dan metode serangan, entitas bisnis memerlukan lebih banyak langkah dan usaha yang dapat melindungi diri mereka sendiri dan data sensitif mereka. Banyak dari serangan siber yang paling sukses baru-baru ini menimbulkan ide tentang saling berbagi *threat intelligence* untuk menggagalkan dan mitigasi dari serangan. Pemerintah dan pelaku bisnis ingin memberikan cara proaktif untuk mempertahankan diri dari serangan siber dengan membagikan informasi ancaman secara tepat waktu [FEN21]. *Intelligence* mengacu pada proses mengumpulkan, menganalisis, dan menafsirkan informasi taktis untuk membuat keputusan. Oleh karena itu, untuk dapat didefinisikan sebagai *intelligence*, informasi harus digabungkan, dianalisis, ditafsirkan, dan disebarluaskan. Informasi mentah, di sisi lain, dapat diperoleh dari segala macam sumber dan dapat menyesatkan, tidak akurat, terputus-putus, dan tidak dapat diandalkan.

Untuk membantu mengurangi serangan siber, perusahaan seperti FireEye dan Cyveillance menyediakan laporan *cyber threat intelligence* (CTI) yang dirancang untuk membantu organisasi melindungi dari serangan siber. Untuk membuat laporan mereka, perusahaan-perusahaan ini mengandalkan data yang dikumpulkan dari serangan atau peristiwa aktual melalui mekanisme seperti log jaringan, log antivirus, *honeypots*, *database access events*, upaya login sistem, dan *intrusion defense system/intrusion protection system* (IDS/IPS) *event log* [SHA15].

The SANS Institute memberikan definisi untuk *cyber threat intelligence* sebagai *threat intelligence* yang berhubungan dengan komputer, jaringan dan teknologi informasi [FEN21]. CTI mengandung berbagai atribut yang dapat menyajikannya sebagai informasi intelijen. Alamat IP atau hash berbahaya sendiri belum dapat dianggap sebagai CTI, tetapi mereka dapat menjadi bagian darinya. Atribut dapat mencakup deskripsi *threat actor*, manuver, motivasi, dan *Indicator of Compromise* (IoC) yang dapat dibagikan kepada para pemangku kepentingan terpercaya [WAG19]. IoC adalah salah satu atribut CTI yang paling mudah

ditindaklanjuti dan merupakan fokus dari sebagian besar model CTI [FAR13]. IoC CTI yang dapat ditindaklanjuti biasanya akan diterapkan dalam aplikasi seperti *Intrusion Detection Systems* (IDS), pemblokiran situs web, *blackholing*, mengidentifikasi *host* dan *malware* yang disusupi [CIO14]. Teknologi *Big Data* digunakan untuk menyimpan atribut CTI dan terdiri atas keterhubungan antara indikator historis dengan indikator yang baru [CHI15]. Atribut CTI lebih berfokus pada TI perusahaan dan cenderung mengabaikan bidang baru seperti *Internet of Things* (IoT), *Industrial Internet of Things* (IIoT) dan area otomotif. Namun demikian, teknologi-teknologi ini, atau lebih sering disebut sebagai *embedded system*, terhubung ke bagian back-end dan dapat mengambil manfaat dan analisis dari atribut CTI yang ditujukan untuk TI perusahaan.

### II.3 Model CTI

Jaringan perusahaan biasanya akan dilengkapi dengan beberapa perangkat keamanan seperti *firewall* tradisional, IDS, IPS, perangkat lunak *anti-malware*, *traffic sniffer*, dll. Penerapan alat-alat ini merupakan investasi dalam keamanan perusahaan, terutama kemampuan untuk melindungi aset perusahaan dan informasi sensitif. Sebagian besar alat ini adalah sistem deteksi berbasis aturan yang mengizinkan atau menolak lalu lintas data sesuai dengan seperangkat aturan yang harus ditentukan sebelumnya. Di sisi lain, keamanan siber adalah sebuah proses dan *life-cycle*, yang membutuhkan improvisasi berkelanjutan. Oleh karena itu, penting untuk berpikir lebih analitis tentang bagaimana menghadapi ancaman siber tingkat lanjut. Aktif mencari ancaman siber adalah pendekatan yang lebih maju dan kompleks daripada sistem deteksi berbasis aturan tradisional. Informasi ancaman siber dapat diakses dari mana saja, termasuk semua rentang waktu di mana ancaman terjadi. Pada model dibawah, metode dibagi berdasarkan *event time of view* [ALM18].

### II.3.1 Post-Event CTI Sharing

*Cyber threat analysis* adalah kunci untuk melakukan threat hunting. *Cyber threat hunting* adalah proses pencarian potensi ancaman siber melalui jaringan dengan menganalisis kumpulan data yang relevan. Analisis data dapat dilakukan dengan otomatis menggunakan alat yang ada atau sebagai alternatif dilakukan secara manual. Dalam sebuah organisasi, kematangan *cyber threat hunting* bergantung pada kemampuan pengumpulan dan analisis data [---16). Data dapat berupa data historis atau *real-time*, tergantung pada sumber yang dipilih untuk mengidentifikasi ancaman siber. Data ancaman dapat dikumpulkan dengan menggunakan *honeypots* dan dianalisis untuk memahami ancaman sebelum terjadi [SON11]. Data juga berisi rincian insiden keamanan siber yang telah terjadi. Menganalisis data tersebut memberikan indikasi bahwa sebagian besar insiden keamanan tidak terjadi sebagai *zero-day attack* [POR06], mereka cukup sering dan dalam banyak kasus memiliki pola. Pengumpulan dan analisis data yang tepat dapat menghasilkan banyak elemen IoC.

Intelijen yang diberikan bersifat reaktif daripada proaktif, karena didasarkan pada data dari serangan yang telah terjadi. Laporan tersebut belum bisa memberikan informasi intelijen yang komprehensif mengenai alat yang telah dikembangkan oleh peretas dan potensi kerusakan yang besar ketika digunakan untuk serangan siber (misalnya, *zero-day attack*). Selain itu, laporan ini sering mengabaikan aktor tertentu yang bertanggung jawab atas eksploitasi semacam itu, sehingga menghasilkan gambaran yang tidak lengkap tentang lingkungan sudut pandang peretas secara keseluruhan.

### II.3.2 Pre-Event CTI Sharing

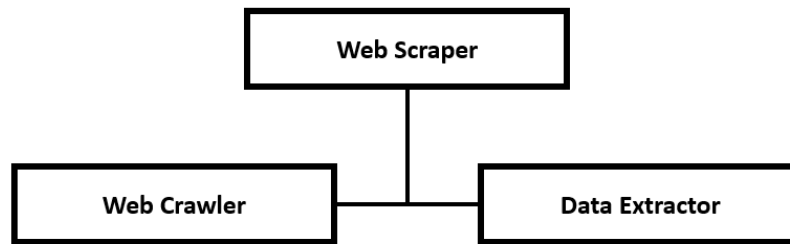
*Pre-reconnaissance cyber threat intelligence* mengacu pada informasi yang dikumpulkan sebelum pihak jahat berinteraksi dengan sistem komputer yang ditinjau. Banyak individu di balik operasi serangan siber, yang berasal dari luar laboratorium pendidikan maupun fasilitas militer yang dikelola pemerintah, bergantung pada komunitas dan forum diskusi yang signifikan. Mereka tetap dapat

berinteraksi melalui berbagai asosiasi *online*, tetap anonim, dan menjangkau rekan-rekannya yang tersebar secara geografis [NUN16]. Nunes, pada penelitiannya, berhasil mengembangkan dan menerapkan sistem pengumpulan data intelijen terkait aktivitas aktor berbahaya. Sistem mereka berhasil beroperasi dan sedang dalam proses pengimplementasian sistem ini ke mitra komersial. Menurutnya, masih banyak tantangan desain untuk mengembangkan *crawler* terfokus menggunakan *data mining* dan teknik *machine learning*. Basis data yang dibangun tersedia bagi para praktisi keamanan untuk mengidentifikasi ancaman dan kemampuan siber yang muncul.

## II.4 Ekstraksi Informasi

World Wide Web adalah ruang informasi global yang berisi jutaan data yang dapat diakses melalui Internet, dan mengekstraksi sejumlah besar data dari web dikenal sebagai *web data extraction* atau *web scraper*. Seperti yang dijelaskan oleh Marres dan Weltevrede pada tahun 2013, dalam jurnalisme, *web scraping* juga telah digunakan untuk menilai signifikansi berita internasional dengan menghitung berapa kali berita tersebut disebutkan oleh pengguna media sosial [MAR13]. Riset di bidang ekstraksi informasi, yang menjelaskan bagaimana konten tidak terstruktur atau semi terstruktur dapat diproses untuk memenuhi tujuan akhir informasi yang ditentukan, telah memungkinkan jenis pemrosesan ini.

Pada tahun-tahun awal, teknik web data scraping yang ada adalah manual *human-copy-paste*. Karena sifat dinamis dan perkembangan teknologi dari dunia web, metode tradisional seperti manual *human-copy-paste* tidak efektif dilakukan. Oleh karena itu, aspek otomatisasi digunakan dalam proses web scraping. Karena bahasa pemrograman yang digunakan untuk menampilkan halaman web modern, yang dikenal sebagai *Hypertext Markup Language* (HTML), terstruktur secara hierarkis menjelaskan makna teks atau konten lain yang dikandungnya, atau disebut sebagai web semantik, ekstraksi data otomatis dari web dapat dibuat. Web scraper secara umum terdiri dari dua bagian, yang pertama adalah *crawler* dan yang kedua adalah ekstraktor data yang ditunjukkan pada Gambar 2.



**Gambar 2 Web Scraper Tree Diagram**

*Web crawler*, *crawler* atau *web spider*, adalah program komputer yang digunakan untuk mencari dan secara otomatis mengindeks konten situs web dan informasi lainnya melalui internet. Program ini paling sering digunakan untuk membuat entri untuk indeks pada search engine. Studi membuktikan bahwa algoritma terbaik untuk perayapan web adalah *Genetic Algorithm*. Untuk mengekstrak data besar dan tidak terstruktur dari web ada beberapa teknik dan alat untuk *data extraction* yang dapat dengan mudah mengekstrak dan mengubahnya menjadi format yang bermakna dan terstruktur. Berikut merupakan beberapa cara ekstraksi data [PAR18].

- a. *Human copy and paste*
- b. *HTML parser*
- c. *HTTP programming*
- d. *Tree-based technique*
- e. *Web Wrapper*

Menurut penelitian yang dilakukan oleh Ferrara pada tahun 2014, secara implisit, *web wrapper* merupakan teknik terbaik untuk mengekstraksi data dari web [FER14].

## **II.5 Forum Internet**

Forum internet adalah papan pesan online yang dibuat agar anggota dapat berkomunikasi satu sama lain dengan mendiskusikan berbagai topik. Meskipun ada

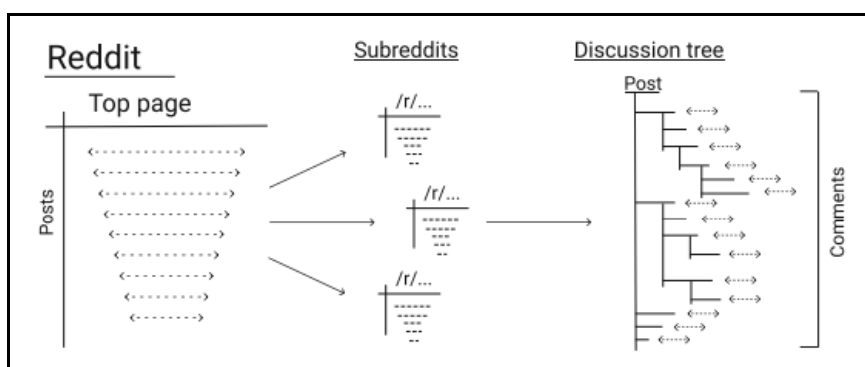
banyak kerangka struktur forum, hampir semua forum memiliki struktur tree yang serupa. Sebuah forum dapat mencakup beberapa subforum, yang masing-masing berfokus pada topik yang berbeda [WIL18]. Pengguna dapat membuat posting atau diskusi yang relevan dengan subtopik tertentu dalam subforum. Dengan memposting, pengguna terlibat dalam percakapan di sebuah utas. Setiap posting menyertakan nama pengguna poster, tanggal posting, dan teks posting. Forum secara otomatis mengarsipkan semua posting; kecuali moderator forum atau poster asli menghapusnya, mereka selalu dapat diakses. Forum adalah *platform* berbasis HTML, sama seperti situs web lainnya. Oleh karena itu, *web crawler* dapat digunakan untuk menjelajahnya dan mengumpulkan konten-kontennya. Alat perangkat lunak yang dikenal sebagai web crawler menjelajahi internet dengan mengklik tautan dan mengumpulkan halaman web menggunakan protokol HTTP [FU10]. *Web crawler* tradisional mengadopsi strategi *Breadth-First Search* (BFS) untuk menavigasi *hyperlink*. Akan tetapi, karakteristik dan struktur forum yang unik membuat *web crawler* biasa yang menggunakan strategi ini tidak mempunyai efisiensi yang baik untuk pengumpulan data [JIA14]. Struktur pohon forum sering menghasilkan halaman/tautan duplikat, halaman tidak informatif, dan tautan membalik halaman. Untuk mengatasi masalah ini, studi sebelumnya terutama menggunakan penguraian dan analitik URL yang dikombinasikan dengan ekspresi reguler untuk menargetkan area tertentu dari forum web untuk pengumpulan [PAV13].

## II.6 Reddit

Diskusi di forum Reddit pada umumnya bersifat publik karena siapa pun, dengan atau tanpa akun Reddit, dapat melihat konten (kecuali *private* subreddit). Visibilitas konten Reddit yang dibagikan maupun komentar suatu diskusi ditentukan oleh “*voting*” pengguna Reddit. Untuk menjadi pengguna Reddit, yang dibutuhkan calon pengguna hanyalah memilih nama pengguna yang unik dan kata sandi. Verifikasi email tidak diperlukan untuk membuat akun. Akan tetapi, *terms of service* Reddit mendikte pengguna harus berusia minimal 13 tahun untuk



mendaftar. Norma umum di sebagian besar situs cenderung menghindari partisipasi dengan nama asli sebagai tindakan perlindungan privasi. Riwayat partisipasi di situs ini juga bersifat publik, artinya siapapun dapat melihat semua komentar dan kiriman publik pengguna dengan mengklik nama pengguna mereka. Kemudahan yang dapat digunakan pengguna untuk membuat akun memungkinkan, dan tidak jarang, satu orang memiliki banyak akun. Akun “*throwaway*”, atau akun “*dummy*” yang dibuat untuk penggunaan waktu terbatas untuk satu tujuan tertentu, biasanya digunakan saat pengguna tidak ingin postingan atau komentar dikaitkan dengan akun utama mereka, seperti berbagi informasi sensitif atau pribadi [AMM19]. Karena partisipasi di Reddit bersifat *pseudonymous*, informasi demografis agak sulit diperoleh. Menurut administrator situs Reddit [RED21] mayoritas (58%) pengguna berusia antara 18 dan 34 tahun dan berjenis kelamin laki-laki (57%).



**Gambar 3 Diagram Skematik Struktur Konten pada Reddit [MED19]**

Gambar 3 menunjukkan struktur konten yang ada di dalam Reddit. *User* yang terdaftar dapat mengirimkan *posting* atau *submission* yang berisi judul, tautan eksternal, atau konten yang ditulis sendiri, yang akan langsung tersedia untuk seluruh pengguna Reddit untuk dapat melakukan *voting* dan berkomentar [MED19]. Sistem voting hanya mengizinkan pengguna terdaftar untuk *upvote* (memberikan suara positif +1) atau *downvote* (memberikan suara negatif -1) pada postingan atau komentar. Komentar membentuk *discussion tree*, yang dapat digambarkan sebagai akar bercabang, di mana simpul akar utama adalah yang

mewakili post itu sendiri dan setiap simpul cabang dibawahnya mewakili komentar. Terdapat hubungan antara dua simpul jika ada hubungan “*reply-to*” di antara keduanya. Ruang *posting* utama Reddit dibagi menjadi subreddit yang berisikan komunitas pengguna yang dibuat sendiri dan disatukan oleh topik tertentu. Setiap kiriman yang dikirim memiliki nama subreddit sebagai atribut yang tersirat. Setiap subreddit dan Reddit sendiri memiliki apa yang disebut “*top page*”, sebagai *feed timeline* tempat judul postingan dengan tautan *voting* dan komentar dikirim ke user. Ada dua faktor yang mempengaruhi posisi peringkat postingan, yaitu waktu dan skor *voting*, atau disebut “karma”, yang pada dasarnya adalah perbedaan antara *upvote* dan *downvote*. Posting skor tinggi memiliki peluang lebih tinggi untuk muncul di halaman atas.

Subreddit dibuat oleh pengguna dan dimoderasi oleh pengguna. Meskipun ada beberapa aturan Reddit yang menyeluruh tentang konten, subreddit sangat bervariasi mengenai konten yang diizinkan, dan dengan konteks ataupun norma khusus pada bahasan tersebut [CHA18]. Sebagai bagian dari aturan khusus subreddit mereka, beberapa subreddit memberikan peringatan kepada peneliti tentang pengumpulan data di komunitas. Misalnya, r/depression dan r/SuicideWatch menyatakan semua *posting* dan survei untuk penelitian harus disetujui oleh tim moderator, atau r/IndianCountry yang melarang penelitian tanpa izin dan meminta siapa pun yang tertarik menggunakan subreddit untuk tujuan penelitian harus melengkapi formulir untuk ulasan oleh moderator.

Selain aturan subreddit secara individual, Reddit juga memiliki *user agreement* di seluruh situs. *User agreement* Reddit [RED20] mencakup pernyataan sebagai berikut terkait pengumpulan data:

“Access, search, or collect data from the Services by any means [automated or otherwise] except as permitted in these Terms or in a separate agreement with Reddit. We conditionally grant permission to crawl the Services in accordance with the parameters set forth in our

*robots.txt file, but scraping the Services without Reddit's prior consent is prohibited."*

Istilah-istilah ini cukup standar dalam ambiguitasnya [FIE20], tetapi pernyataan ini menyiratkan bahwa data yang dikumpulkan di luar batas kelonggaran tertentu dapat dilakukan, misalnya, menggunakan *Application Programming Interface* (API) Reddit. Hal tersebut mungkin merupakan pelanggaran terhadap *user agreement* ini. Namun, API Reddit tersedia secara gratis dan dapat digunakan untuk mengakses konten di situs Reddit.

## **II.7 Penelitian Terkait**

### **II.7.1 Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics [PRO21]**

Tujuan dari penelitian ini adalah untuk menjelaskan bagaimana Reddit digunakan oleh para peneliti sebagai sumber data. Pertama, penelitian ini mencatat peningkatan jumlah studi yang menggunakan data Reddit selama sepuluh tahun sebelumnya. Sebagian besar proses dilakukan untuk menghasilkan artikel penelitian yaitu dengan menggunakan teknik komputasi seperti web scraping. Topik yang dipelajari menggunakan data Reddit sangat bervariasi, dan terkadang, peneliti mengambil data dari komunitas di Reddit yang mungkin mencakup populasi yang rentan. Kesimpulannya, hanya sedikit peneliti yang membagikan ilmu yang mereka hasilkan di Reddit, namun hampir 30% penelitian mengenai Reddit di data artikel penelitian ini muncul di Reddit. Ini menunjukkan adanya minat pada Reddit secara luas untuk penelitian tentang Reddit. Namun, eksplorasi lebih lanjut diperlukan untuk lebih memahami nilai yang diciptakan oleh keterlibatan dan berbagi pengetahuan semacam ini.

### **II.7.2 Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence [SAM17]**

Penelitian ini mengembangkan framework baru untuk CTI dengan memanfaatkan pendekatan mining baik web, data, maupun teks secara otomatis dan

berprinsip untuk mengumpulkan dan menganalisis sejumlah besar *source code hacker*, *tutorial*, dan lampiran langsung dari komunitas internasional hacker bawah tanah yang besar. *Framework* ini memungkinkan peneliti untuk mengidentifikasi banyak aset jahat yang tersedia secara bebas di forum hacker bawah tanah seperti *crypters*, *keyloggers*, *SQL Injections*, dan *cracker password*, beberapa di antaranya mungkin menjadi akar penyebab pelanggaran baru-baru ini terhadap organisasi seperti pada Kantor Manajemen Personalia Amerika Serikat. Peneliti juga dapat menentukan individu utama di balik aset ini dengan menggunakan teknik dan metrik analisis jaringan sosial. Pendekatan dapat digeneralisasikan ke forum peretas mana pun, terlepas dari struktur subforum. Penelitian ini memiliki implikasi praktis bagi organisasi yang ingin meningkatkan postur keamanan siber mereka. Dengan asumsi sebuah organisasi mengetahui sistem yang ingin dilindunginya, mereka dapat menerapkan kerangka kerja ini ke forum yang dipilihnya untuk mengidentifikasi aset peretas yang relevan untuk sistem mereka.

## **BAB III**

### **DESAIN DAN IMPLEMENTASI**

Bab ini menjelaskan mengenai analisis permasalahan yang akan dikaji berupa analisis mengenai hambatan-hambatan dari sistem yang sudah dikembangkan. Selain itu, dilakukan juga penjelasan mengenai solusi yang ditawarkan berupa konsep solusi yang akan direalisasikan. Terakhir, akan dijelaskan mengenai timeline penyelesaian tugas akhir.

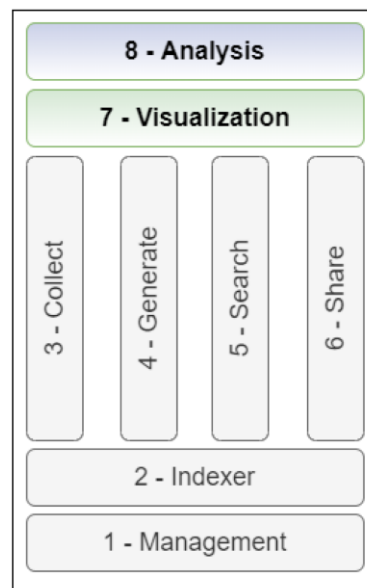
#### **III.1 Analisis Permasalahan**

*Web scraping* sering kali menyerempet *terms of service* yang tertera situs web target. *Terms of service* situs-situs berbasis data yang berat hampir selalu melarang *data scraping* secara ilegal. Melanggar *terms of service* bukan berarti telah melakukan sesuatu yang mengancam diri sendiri. Akan tetapi, pihak situs web tersebut mungkin saja dapat menuntut atas pelanggaran kontrak dan dapat dibawa ke ranah hukum. Terlebih lagi, sebagian besar situs web yang tidak ingin dilakukan scraping mempunyai metode penanggulangannya. Web-web itu hanya ingin menyajikan konten kepada pengguna manusia asli yang menggunakan *browser* web, dengan pengecualian jika menyangkut *crawler* milik Google agar terlihat pada pencarian Google. Oleh karena, saat dilakukan *web scraping*, sistem akan menganggap akses dilakukan oleh robot, sehingga terdapat pembatasan. Hal itu terjadi karena sistem tidak menemukan ciri-ciri manusia pada pengaksesannya. Terdapat dua faktor utama untuk dikenali sebagai manusia, yaitu menggunakan peralatan milik manusia (*browser*) atau memiliki perilaku seperti manusia.

Reddit merupakan salah satu platform forum diskusi yang sering digunakan untuk membicarakan berbagai topik. Akan tetapi, terdapat satu permasalahan tentang Reddit yang cukup trivial dan menyangkut hal yang bersifat nonteknis. Reddit merupakan salah satu yang termasuk daftar situs yang dilarang oleh Pemerintah Indonesia. Hal ini mengakibatkan akses ke situs tersebut dibatasi secara

umum. *Proxy* maupun *Virtual Private Network* (VPN) dibutuhkan untuk dapat mengakses Reddit dengan internet publik di wilayah Indonesia.

Terdapat beberapa data points di Reddit yang dapat dilakukan ekstraksi data menggunakan web scraping. *Data points* ataupun sumber data tersebut adalah subreddit, *submission* pada sebuah subreddit, konten yang ada di dalam *post* atau *submission*, seperti: halaman, tautan, komentar, gambar, *upvotes* dan *downvotes*. Selain itu, juga terdapat banyak sumber data lainnya sesuai dengan kebutuhan. Reddit merupakan salah satu *website* yang menyediakan *open source* API untuk beberapa keperluan seperti bisnis dan penelitian. API Reddit dapat dimanfaatkan untuk melakukan *scraping* dengan langsung mengirimkan data dari Reddit ke target tanpa harus mengakses *website* Reddit. Akan tetapi, ada pembatasan bahwa Reddit tidak dapat menyediakan konten untuk dilakukan *scraping* selain 1000 konten teratas atau terpopuler menggunakan Reddit API.



**Gambar 4 CTI 8-step Model [AMA22]**

Kebutuhan pada sistem CTI dapat dijelaskan menggunakan CTI *8-step model* yang tertera pada Gambar 4 [AMA22]. *Step 1 - Management* bertanggung jawab untuk mengelola interaksi pengguna dengan setiap fungsionalitas yang ditawarkan aplikasi dan interaksi di antara mereka. Step ini memiliki kebutuhan untuk mengontrol aliran data dan aksesnya melalui pengelolaan izin akses. *Step 2 - Indexer* didukung oleh struktur penyimpanan yang mampu mendukung jumlah input dan output data pada sistem CTI. *Step 3 - Collect* bertanggung jawab untuk menyediakan pengumpulan dan penyisipan data eksternal ke dalam sistem CTI. *Step - 4 Generate* dapat menyediakan normalisasi data internal menggunakan pola *Step 2* dan *3* yang sama sehingga data yang diserap dapat diubah menjadi *feeds* untuk digunakan nanti di sistem CTI. *Step 5 - Search* dapat menyediakan mekanisme dan metode untuk memungkinkan manipulasi dan eksplorasi data secara efektif serta memungkinkan pengindeksan data yang cepat dan visibilitas penuh data yang disimpan. *Step 6 - Share* dapat mengizinkan berbagi data internal seperti *feeds*, koleksi, dan indikator ancaman antar pengguna, serta berbagi dengan alat pihak ketiga. *Step 7 - Visualization* dapat memberikan visualisasi data CTI dalam format temporal untuk membuat indikator ancaman sehingga seseorang dapat memiliki gambaran lengkap dari jejak ancaman, IoC, dan informasi berguna apa pun yang diperkaya dan dibagikan oleh pihak lain yang berkepentingan. *Step 8 - Analysis*, yang merupakan bagian intrinsik dari *Step 7*, dapat mengimplementasikan fungsionalitas yang memungkinkan analisis data dan memanipulasi serta memperoleh informasi terbaik dari data ancaman yang tersedia.

Untuk memenuhi *Step 7 - Visualization*, CTI harus dapat menampilkan informasi dan *insight* yang dapat dipahami oleh pihak terkait. Informasi maupun *insight* ini dapat berbentuk sebuah peringatan maupun saran terkait keamanan siber dari sistem yang ditinjau. Akan tetapi, hasil dari *web scraping* masih dalam bentuk *raw data* atau *data frame*. Perlu pengolahan lebih lanjut dari kumpulan data tersebut menjadi informasi yang berharga.

Analisis kebutuhan sistem diperlukan untuk memahami permasalahan dengan melihat gambaran awal dari sistem dan apa saja yang dapat dilakukannya.

Analisis kebutuhan dapat berupa fungsional dan non fungsional. Tabel 1 dan Tabel 2 menunjukkan analisis kebutuhan fungsional dan non fungsional.

**Tabel 1 Kebutuhan Fungsional Sistem**

ID	Kebutuhan	Deskripsi
FR-1	Sistem dapat mengirimkan API Request ke Reddit	Sistem mengirimkan metode GET ke API Reddit untuk mengambil data terkait konten yang ada di subreddit sesuai dengan tautan yang sudah ditentukan.
FR-2	Sistem dapat menyaring data yang diperlukan	Data yang didapat dari respon API Reddit masih bersifat umum dan berisi semua metadata dan segala isinya. Sistem dapat memilih data berdasarkan parameter yang diinginkan untuk mempermudah pengolahan.
FR-3	Sistem dapat memberikan visualisasi terhadap data yang berhasil didapatkan	Visualisasi dilakukan untuk menyajikan data dalam bentuk yang mudah dipahami sehingga akan mempermudah analisis. Oleh karena itu, sistem dapat menampilkan data yang sudah diubah ke dalam bentuk grafis.
FR-4	Sistem dapat memberikan tanda bahwa subreddit yang dianalisis mempunyai level ancaman tertentu	Sistem dapat memberikan flag kepada suatu konten hasil dari Reddit <i>Web Scraping</i> . Dengan adanya <i>flag</i> tersebut, CTI dapat memberikan makna kepada pihak terkait untuk memberikan perhatian khusus ke konten tersebut.



**Tabel 2 Kebutuhan Non Fungsional Sistem**

ID	Kebutuhan	Deskripsi
NFR-1	95% <i>response time</i> dari pemanggilan API Reddit tidak lebih dari 500 ms	Rentang waktu yang diberikan ketika mengirim API GET <i>request</i> ke Reddit dan sistem menerima data tidak melebihi 500 ms dalam 95% dari semua kesempatan.
NFR-2	Sistem dapat menyimpan data hasil <i>crawling</i> minimal 1000 data per satu <i>cycle</i>	Data yang disimpan setiap satu <i>cycle web crawling</i> minimal berjumlah 1000 buah. Data yang disimpan disesuaikan dengan aturan Reddit yang hanya mengizinkan maksimal 1000 konten untuk dilakukan <i>scraping</i> .
NFR-3	Seluruh level akses sistem dapat digunakan oleh pengguna	Sistem dapat digunakan sepenuhnya oleh pengguna. Untuk saat ini, level akses belum diberikan spesialisasi karena sistem ini memiliki satu fungsi utama sehingga setiap penggunaan memiliki tujuan yang sama.

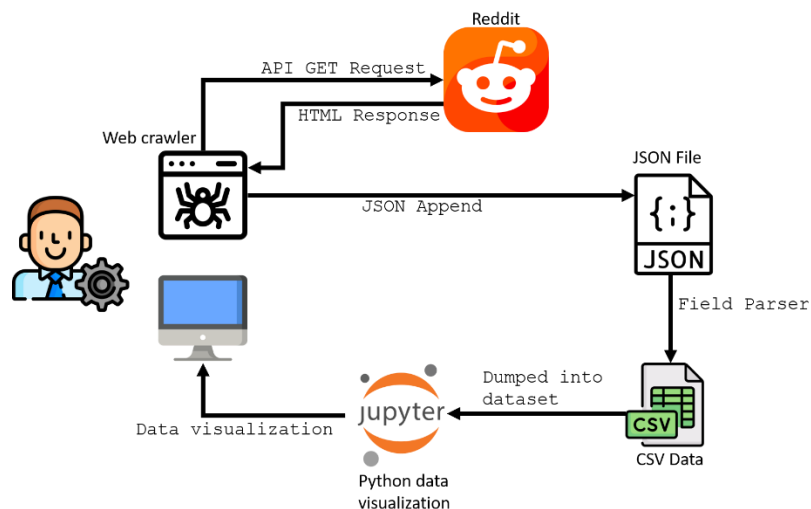
### **III.2 Rancangan Solusi Secara Garis Besar**

Ada beberapa pendekatan untuk melakukan *web scraping* pada Reddit yang dibagi berdasarkan level keterlibatan pengguna. Pendekatan pertama ialah *scraping* secara manual. *Scraping Reddit* secara manual adalah cara yang termudah tetapi paling tidak efisien dalam hal kecepatan dan biaya. Akan tetapi, *scraping* secara manual menghasilkan data dengan konsistensi tinggi. *Scraping* secara manual

cocok untuk kebutuhan scraping yang terbatas hanya untuk beberapa utas Reddit tentang topik tertentu. Pendekatan kedua adalah scraping menggunakan Reddit API. Cara ini dapat menghasilkan data dengan mudah tetapi untuk menjalankannya, diperlukan setidaknya keterampilan dan kompetensi pemrograman. Selain itu, Reddit API membatasi jumlah postingan di *data point* mana pun maksimal 1000 buah data. Pendekatan ketiga adalah dengan memanfaatkan layanan API pihak ketiga untuk scraping Reddit. Cara ini adalah pendekatan yang efektif dan mempunyai skalabilitas yang tinggi tetapi menggunakan biaya yang besar. Cara ini hanya dilakukan apabila kebutuhan data dari scraping Reddit melewati beberapa juta postingan atau konten.

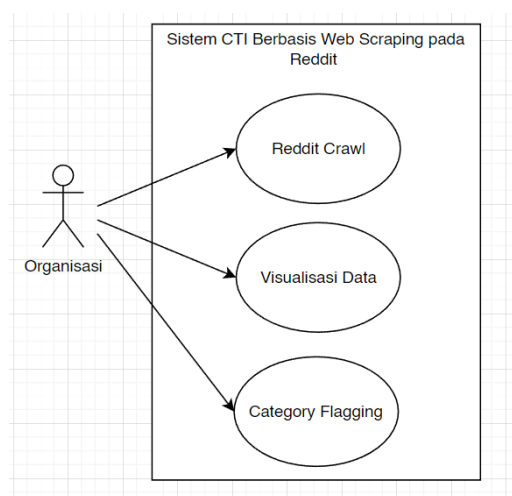
Pembangunan dan implementasi *Reddit Scraper* pada penelitian ini menerapkan pendekatan kedua yakni memanfaatkan API Reddit yang telah disediakan. Selain dapat disesuaikan dengan topik yang akan dilakukan *scraping*, pendekatan ini juga tidak memerlukan server untuk mesin *crawler* ditempatkan. Mesin *crawler* dapat dijalankan di komputer lokal sehingga dapat menghindari pemakaian API Reddit yang tidak wajar dan menghindari penyalahgunaan *terms of service* pada Reddit sendiri.

Untuk menggambarkan keseluruhan sistem *Web Scraping* pada Reddit, digambarkan diagram arsitektur seperti pada Gambar 5.



**Gambar 5 Diagram Arsitektur Sistem CTI Berbasis *Web Scraping* pada Reddit**

Gambar 5 menunjukkan arsitektur sistem CTI berbasis web scraper pada Reddit. Sistem bekerja dengan *web crawler* yang bekerja aktif memanggil API endpoint Reddit. *Web crawler* dibangun menggunakan bahasa python dan mengimplementasikan *library* khusus bernama PRAW (Python Reddit API Wrapper). Library ini memungkinkan program untuk berinteraksi dengan Reddit melalui Python. *Web crawler* akan mengirimkan sebuah request GET ke Reddit, lalu Reddit akan merespon dengan mengirimkan konten sesuai dengan tautan subreddit yang tertulis di program. Konten yang dikirimkan dari Reddit masih terdiri dari keseluruhan metadata setiap jenisnya. Untuk visualisasi CTI, hanya diperlukan data-data yang dianggap esensial seperti isi komentar dan nama *user*. Oleh karena itu, dari keseluruhan metadata, hanya diambil beberapa untuk dilakukan dumping untuk mendapatkan dataset. Dataset tersebut dapat divisualisasikan menggunakan python *data visualization* dan disajikan ke pihak yang berkepentingan.



**Gambar 6 Diagram Use Case Sistem CTI Berbasis Web Scraping pada Reddit**

Gambar 6 merupakan *use case diagram* dari sistem. Sistem mempunyai tiga fungsi utama yaitu Reddit Crawl untuk mengambil data langsung dari Reddit dan Visualisasi Data untuk membuat data dari Reddit dapat dipahami dengan mudah. Pengkategorian tipe ancaman dilakukan juga ketika Visualisasi Data. Data-data

yang sudah dikelompokkan saat visualisasi akan diberikan *flag* untuk kata kunci yang sering muncul. Kata-kata yang dipilih sebagai kata kunci adalah kata-kata yang berpotensi mengancam keamanan seperti “*data breach*”, “*password leak*”, “*harass*” dan lain-lain. Jika pada suatu subreddit memiliki banyak *flag* yang sudah dijelaskan di atas, maka akan diberikan nilai dengan skala tertentu berbanding lurus dengan jumlah *flag*. Semakin tinggi skala maka subreddit akan semakin ditandai sebagai ancaman.

### III.3 Linimasa Penyelesaian Tugas Akhir

Tabel 3 berikut merupakan timeline penyelesaian tugas akhir berdasarkan bab yang akan dikerjakan.

**Tabel 3 Timeline Penyelesaian Tugas Akhir**

Pekerjaan	2022				2023			
	September	Oktober	November	Desember	Januari	Februari	Maret	April
Bab 1 Pendahuluan								
Bab 2 Dasar Teori								
Bab 3 Desain dan Implementasi								
Bab 4 Pembahasan								

Bab 5 Kesimpulan dan Saran								
----------------------------------	--	--	--	--	--	--	--	--

## BAB IV

### IMPLEMENTASI DAN EVALUASI

Tujuan penulisan bab ini adalah untuk menunjukkan seberapa jauh solusi yang diuraikan pada bagian sebelumnya dapat menyelesaikan permasalahan utama Tugas Akhir. Metode yang dipakai adalah pengujian berdasarkan skenario yang dibangun untuk memvalidasikan kebutuhan yang sudah dituliskan di bab sebelumnya.


#### IV.1 Ekstraksi Data

Untuk mendapatkan dataset yang nantinya akan dipergunakan untuk analisis dan visualisasi, data diambil secara terpisah dari setiap komentar dan *submission* di dalam Reddit. Data yang diambil secara terpisah tersebut, selanjutnya akan dikumpulkan dan digabungkan dalam satu file tertentu sehingga dapat divisualisasikan dan dianalisis.

##### IV.1.1 Reddit Developer API

Ekstraksi data di Reddit menggunakan layanan yang diberikan oleh Reddit berbentuk API. Untuk dapat menggunakan layanan tersebut, pengguna Reddit harus membuat aplikasi di Reddit Developer sehingga Reddit dapat memberikan *personal script* dan *secret code*.

developed applications



The screenshot shows the 'developed applications' page for a user named 'tugas\_akhir'. The application is titled 'Bachelor thesis' and is marked as a 'personal use script'. The secret key is '8-8kRY17k-I0KsHRh7ey-ZCQjks8vA'. The application is currently owned by 'iammedesu (that's you!)'. The interface includes fields for the application name ('tugas\_akhir'), description ('Bachelor thesis'), about URL, and redirect URI ('http://localhost:8080'). There are buttons for 'update app' and 'delete app'.

secret	8-8kRY17k-I0KsHRh7ey-ZCQjks8vA	developers	iammedesu (that's you!) remove
name	tugas_akhir	add developer: <input type="text"/>	
description	Bachelor thesis		
about url	<input type="text"/>		
redirect uri	http://localhost:8080		
<input type="button" value="update app"/> <input type="button" value="delete app"/>			

Gambar X merupakan aplikasi untuk ekstraksi data yang didaftarkan ke Reddit Developer. *Personal use script* dan *secret* digunakan untuk mengakses Reddit API yang diimplementasikan menggunakan kode program Python.

#### IV.1.2 Python Reddit API Wrapper (PRAW)

Python Reddit API Wrapper (PRAW) merupakan paket *library* bahasa pemrograman Python yang memungkinkan akses sederhana ke API Reddit.

##### IV.1.2.1 Konfigurasi PRAW

Untuk memulai dan melakukan modifikasi program menggunakan *library* PRAW, dibutuhkan *instance class* Reddit. Terdapat dua jenis instance yaitu: *read-only* dan *authorized*.

###### IV.1.2.1.1 Read-only Instance

Untuk membuat Reddit read-only instance, dibutuhkan 3 jenis data, yaitu *client ID*, *client secret*, dan *user agent*. Sehingga inisiasi *class* diekspresikan sebagai berikut.

```
reddit ← praw.Reddit {  
  client_id ← "personal use script",  
  client_secret ← "secret code",  
  user_agent ← "user agent",  
}
```

###### IV.1.2.1.2 Authorized Instance

Untuk membuat Reddit *authorized instance*, diperlukan dua informasi tambahan ke dalam inisiasi *class*. Sehingga inisiasi *class* dapat diekspresikan sebagai berikut.

```
reddit ← praw.Reddit {  
  client_id ← "personal use script",  
  client_secret ← "secret code",  
  password ← "password akun Reddit",  
  user_agent ← "user agent",  
  username ← "username akun Reddit",  
}
```

*Authorized instance* memiliki jangkauan yang lebih luas untuk mengakses informasi Reddit dibandingkan *read-only instance* termasuk menuliskan posting ke Reddit. Namun, sistem yang dibangun saat ini, hanya menggunakan *read-only instances* karena hanya memerlukan pengambilan data yang bersifat *public* dari Reddit.

#### **IV.1.2.2 Pengambilan Data Submission dari Subreddit**

Setelah membuat *instance* subreddit, iterasi dapat dilakukan melalui *submission* di dalamnya. Setiap *submission* memiliki *instance* tersendiri yang tersusun dari beberapa data. Beberapa jenis *instance submission* mencakup metode sebagai berikut: *controversial*, *gilded*, *hot*, *new*, *rising*, dan *top*. Masing-masing metode ini akan segera melakukan *return* ListingGenerator, yang akan diiterasi di giliran berikutnya. Misalnya, melakukan iterasi melalui 10 kiriman pertama menggunakan *hot sort* untuk subreddit tertentu.

#### **IV.1.2.3 Pengambilan Data Komentar dari Submission**

Submission memiliki atribut komentar yang merupakan *instance CommentForest*. Instance itu dapat dilakukan iterasi dan merepresentasikan komentar top-level dari submission dengan jenis comment sort default. Jika ingin mengiterasi semua komentar sebagai list yang diratakan, metode *list()* dapat dipanggil pada *instance CommentForest*.

### **IV.1.3 Struktur Data**

Data yang diperoleh dari API Reddit merupakan *instance* dari sebuah gabungan data yang menyusun sebuah item submission maupun komentar. Diperlukan seleksi data-data yang relevan ke dalam satu dataset sehingga memudahkan untuk memahami data tersebut.

#### **IV.1.3.1 Raw Data**

Terdapat data submission dan komentar yang diekstraksi menggunakan API Reddit. Data tersebut memiliki struktur yang mirip. Hubungan antara data submission dan



data komentar adalah one-to-many, sehingga satu submission dapat memiliki beberapa komentar. Hal tersebut dapat dibuktikan bahwa di dalam data komentar terdapat atribut submission yang berfungsi sebagai foreign key.

#### IV.1.3.1.1 Submission

Tabel X menunjukkan atribut apa saja yang terdapat dalam satu item submission dari Reddit API.

Attribute	Description
<code>author</code>	Provides an instance of <code>Redditor</code> .
<code>author_flair_text</code>	The text content of the author's flair, or <code>None</code> if not flaired.
<code>clicked</code>	Whether or not the submission has been clicked by the client.
<code>comments</code>	Provides an instance of <code>CommentForest</code> .
<code>created_utc</code>	Time the submission was created, represented in <u>Unix Time</u> .
<code>distinguished</code>	Whether or not the submission is distinguished.
<code>edited</code>	Whether or not the submission has been edited.
<code>id</code>	ID of the submission.
<code>is_original_content</code>	Whether or not the submission has been set as original content.
<code>is_self</code>	Whether or not the submission is a selfpost (text-only).
<code>link_flair_template_id</code>	The link flair's ID.
<code>link_flair_text</code>	The link flair's text content, or <code>None</code> if not flaired.
<code>locked</code>	Whether or not the submission has been locked.
<code>name</code>	Fullname of the submission.
<code>num_comments</code>	The number of comments on the submission.
<code>over_18</code>	Whether or not the submission has been marked as NSFW.

Attribute	Description
<code>permalink</code>	A permalink for the submission.
<code>poll_data</code>	A <code>PollData</code> object representing the data of this submission, if it is a poll submission.
<code>saved</code>	Whether or not the submission is saved.
<code>score</code>	The number of upvotes for the submission.
<code>selftext</code>	The submissions' selftext - an empty string if a link post.
<code>spoiler</code>	Whether or not the submission has been marked as a spoiler.
<code>stickied</code>	Whether or not the submission is stickied.
<code>subreddit</code>	Provides an instance of <code>Subreddit</code> .
<code>title</code>	The title of the submission.
<code>upvote_ratio</code>	The percentage of upvotes from all votes on the submission.
<code>url</code>	The URL the submission links to, or the permalink if a selfpost.

#### IV.1.3.1.2Komentar

Tabel X menunjukkan atribut apa saja yang terdapat dalam satu item komentar dari Reddit API.

Attribute	Description
<code>author</code>	Provides an instance of <code>Redditor</code> .
<code>body</code>	The body of the comment, as Markdown.
<code>body_html</code>	The body of the comment, as HTML.
<code>created_utc</code>	Time the comment was created, represented in <u>Unix Time</u> .
<code>distinguished</code>	Whether or not the comment is distinguished.

Attribute	Description
<code>edited</code>	Whether or not the comment has been edited.
<code>id</code>	The ID of the comment.
<code>is_submitter</code>	Whether or not the comment author is also the author of the submission.
<code>link_id</code>	The submission ID that the comment belongs to.
<code>parent_id</code>	The ID of the parent comment (prefixed with <code>t1_</code> ). If it is a top-level comment, this returns the submission ID instead (prefixed with <code>t3_</code> ).
<code>permalink</code>	A permalink for the comment. <code>Comment</code> objects from the inbox have a <code>context</code> attribute instead.
<code>replies</code>	Provides an instance of <code>CommentForest</code> .
<code>saved</code>	Whether or not the comment is saved.
<code>score</code>	The number of upvotes for the comment.
<code>stickied</code>	Whether or not the comment is stickied.
<code>submission</code>	Provides an instance of <code>Submission</code> . The submission that the comment belongs to.
<code>subreddit</code>	Provides an instance of <code>Subreddit</code> . The subreddit that the comment belongs to.
<code>subreddit_id</code>	The subreddit ID that the comment belongs to.

#### IV.1.3.2 Parsing Data

Data mentah yang didapat dari submission dan komentar akan dipilih mana saja yang relevan untuk dianalisis. Pemilihan langsung dilakukan dengan program dan otomatis memasukkannya ke dalam file comma separated value (CSV).

##### IV.1.3.2.1 Parsing Data Submission

Berikut merupakan data submission yang diseleksi dan disimpan ke dalam file CSV.

Kolom	Deskripsi
id	Sebagai pembeda antar submission di dalam tabel
created_utc	Waktu submission tersebut dibuat (dalam bentuk Epoch Unix Time)
author	Username pengguna reddit yang membuat submission
num_comments	Jumlah komentar yang diberikan di submission
title	Judul submission
selftext	Self text submission
full_link	URL submission

#### IV.1.3.2.2 Parsing Data Komentar

Berikut merupakan data komentar yang diseleksi dan disimpan ke dalam file CSV.

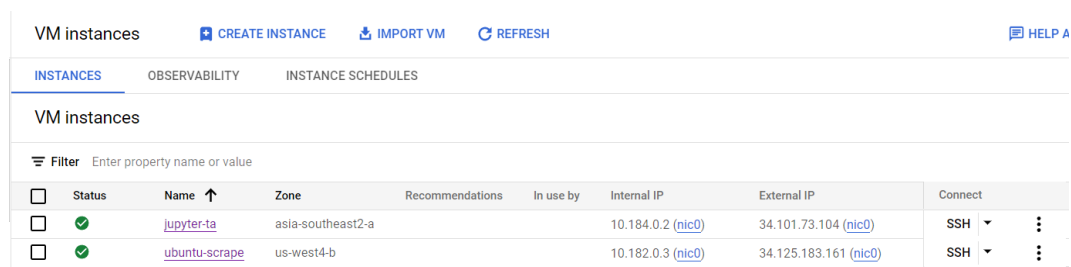
Kolom	Deskripsi
id	Sebagai pembeda antar komentar di dalam tabel
submission_id	ID sebuah submission yang diberikan komentar
created_utc	Waktu komentar tersebut dibuat (dalam bentuk Epoch Unix Time)
author	Username pengguna reddit yang memberikan komentar
score	Akuulasi jumlah upvotes dikurangkan dengan downvotes
body	Isi dari komentar tersebut
parent_id	Menandakan branch komentar tersebut
permalink	URL yang langsung merujuk kepada komentar tersebut

## IV.2 Virtual Machine

Virtual Machine (VM) digunakan untuk menggantikan sistem lokal dalam menjalankan proses crawling/ekstraksi data dari Reddit. Penggunaan VM diharapkan memiliki kinerja yang lebih baik daripada menggunakan sistem lokal dikarenakan memiliki resource yang lebih baik. Selain itu, VM dapat berjalan dalam waktu yang relatif lebih lama, sehingga dapat melakukan ekstraksi data dalam jumlah banyak.

### IV.2.1 Google Cloud Platform (GCP)

Google Cloud Platform (GCP) adalah layanan yang disediakan untuk mendukung operasional perusahaan IT dan pengembang aplikasi. Google Cloud menawarkan layanan untuk komputasi, penyimpanan, jaringan, big data, machine learning, dan IoT, serta pengelolaan cloud, keamanan, dan developer tools. Gambar X merupakan beberapa instans yang berjalan di Google Cloud Platform untuk pengerjaan Tugas Akhir. Terdapat dua instans yang berjalan, masing-masing adalah Ubuntu Virtual Machine dan Jupyter Notebook Server.

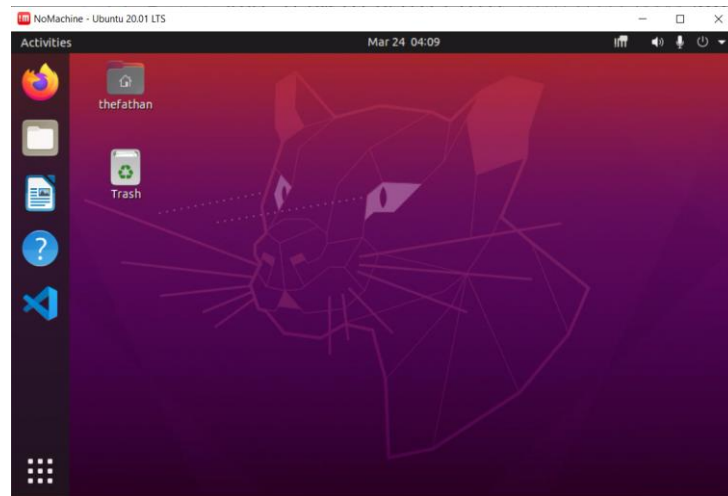


The screenshot shows the 'VM instances' page in the Google Cloud Platform console. At the top, there are buttons for 'CREATE INSTANCE', 'IMPORT VM', 'REFRESH', and 'HELP'. Below these are tabs for 'INSTANCES', 'OBSERVABILITY', and 'INSTANCE SCHEDULES'. The 'INSTANCES' tab is selected, showing a list of VM instances. The list has columns for Status, Name, Zone, Recommendations, In use by, Internal IP, External IP, and Connect. Two instances are listed: 'jupyter-ta' in the 'asia-southeast2-a' zone with internal IP 10.184.0.2 and external IP 34.101.73.104, and 'ubuntu-scrape' in the 'us-west4-b' zone with internal IP 10.182.0.3 and external IP 34.125.183.161. Both instances have a status of 'Running' (indicated by a green checkmark) and an 'SSH' connection option.

Status	Name	Zone	Recommendations	In use by	Internal IP	External IP	Connect
Running	jupyter-ta	asia-southeast2-a			10.184.0.2 (nic0)	34.101.73.104 (nic0)	SSH
Running	ubuntu-scrape	us-west4-b			10.182.0.3 (nic0)	34.125.183.161 (nic0)	SSH

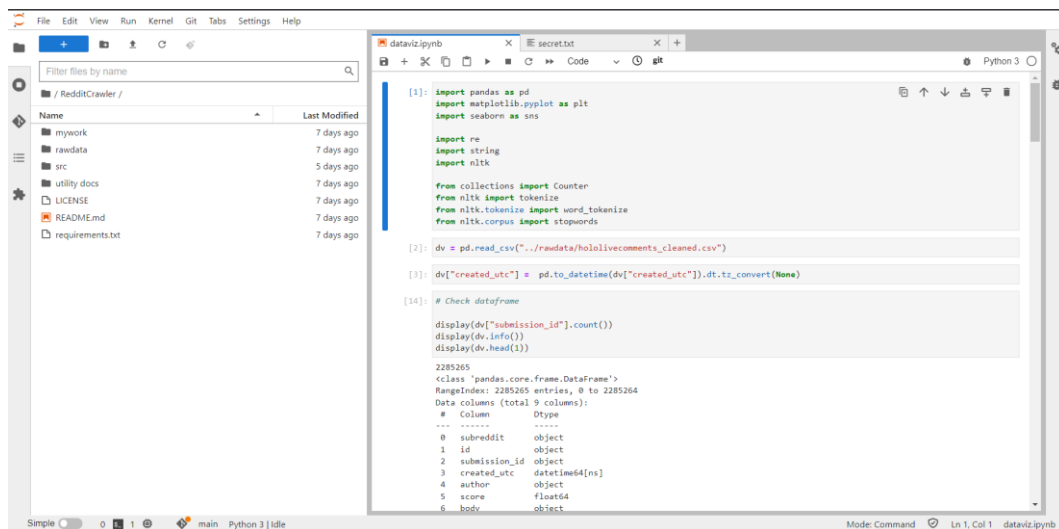
#### IV.2.1.1 Ubuntu Virtual Machine

Virtual machine digunakan untuk menjalankan OS Ubuntu. Proses crawling/ekstraksi data dilakukan menggunakan python environment yang berjalan di atas OS Ubuntu. Penggunaan Ubuntu Virtual Machine dimaksudkan akan proses ekstraksi data yang memerlukan waktu lama dapat terus berjalan di sisi server sehingga tidak diperlukan sistem lokal untuk berjalan secara terus-menerus. Gambar X merupakan Ubuntu Virtual Machine yang digunakan untuk berjalannya proses ekstraksi data pada Tugas Akhir ini.



#### IV.2.1.2 Jupyter Notebook Server

Beberapa proses komputasi data yang berjumlah jutaan memerlukan waktu yang lama untuk mengeksekusinya. Oleh karena itu, diperlukan juga notebook/workbench yang dapat berjalan di sisi server. Google Cloud Platform mempunyai layanan bernama Vertex AI yang dapat difungsikan untuk pengolahan big data dan machine learning. Gambar X merupakan tampilan dari notebook yang berjalan di GCP.



### IV.2.2 NoMachine Remote Desktop

NoMachine adalah aplikasi perangkat lunak lintas platform berpaten untuk akses jarak jauh, berbagi desktop, desktop virtual, dan transfer file antar komputer. NoMachine dapat diinstal pada komputer dengan OS Windows, Mac, Linux, Raspberry Pi dan Linux ARM untuk memungkinkan pengguna mengakses desktop dari jarak jauh melalui jaringan. Pengguna dapat terhubung dari Windows, macOS, iOS, Android, Linux, Raspberry Pi, Linux ARM atau browser web. NoMachine pada Tugas Akhir ini, digunakan untuk menjalankan virtual machine GCP dengan koneksi yang sudah disediakan. Gambar X merupakan tampilan awal NoMachine yang menunjukkan pilihan sistem remote tersimpan.



### IV.3 Pengujian Kemampuan Crawling Sistem

Pengujian ini dilakukan untuk melihat seberapa reliabel sistem dalam mengambil data kasar dari Reddit. Pengambilan data dilakukan dalam beberapa kondisi dan dibandingkan untuk mencari metode yang paling efektif dalam menjalankan skenario tersebut.

#### IV.3.1 Pengujian Lama Waktu Proses Berdasarkan Network

Pengujian ini bertujuan untuk mendapatkan perbandingan lama waktu yang dibutuhkan untuk menyelesaikan satu skenario kondisi yang akan diujikan. Kondisi yang diuji adalah jenis *network* yang digunakan oleh mesin dalam menjalankan

fungsi ekstraksi data dari API Reddit. Pengujian ini menggunakan perintah yang sama, sehingga data yang didapatkan juga merupakan data yang identik. Ekstraksi data dilakukan dengan 10 *batch* dalam 3 *lap*. *Batch* merupakan limit yang ditetapkan dalam mengambil *submission* dalam satu waktu, sedangkan *lap* merupakan berapa kali pemanggilan *batch* dalam satu perintah. Dengan menggunakan nilai *batch* dan *lap* yang sama, diharapkan bahwa item yang didapatkan dari proses ekstraksi juga memiliki jumlah yang sama. Perintah terminal yang digunakan adalah.

```
python src/subreddit_downloader.py Hololive --batch-size 10 --laps 3 --reddit-id DnpX9tZO75idVbdUuDUgdg --reddit-secret 8-8kRY17k-I0KsHRh7ey-ZCQjks8vA --reddit-username iammedesu --utc-before 1676946171
```

Internet Provider	Proxy/VPN	Kecepatan unduh (Mb/detik)	Kecepatan ping (ms)	Waktu rata-rata setiap lap (menit)	Waktu total (menit)	Item yang didapat (row)
Biznet	-	82	11	Gagal	Gagal	Gagal
Biznet	ITB VPN	57	40	0,46	1,4	157
Biznet	Cloudflare 1.1.1.1	88	256	0,86	2,6	157
Firstmedia	ITB VPN	28	23	0,16	0,5	157
Firstmedia	Cloudflare 1.1.1.1	94	266	0,13	0,4	157



Internet Provider	Proxy/VPN	Kecepatan unduh (Mb/detik)	Kecepatan ping (ms)	Waktu rata-rata setiap lap (menit)	Waktu total (menit)	Item yang didapat (row)
Eduroam	-	15	19	0,26	0,8	157
Cloud network	-	1100	8	0,03	0,1	157

Dari beberapa skenario pengujian di Tabel X, didapatkan bahwa proses tercepat untuk perolehan 157 rows data adalah dengan menggunakan cloud network. Dapat pula disimpulkan bahwa cloud network mempunyai kecepatan unduh dan nilai ping yang lebih baik dibandingkan dengan skenario pengujian network lainnya. Oleh karena itu, penggunaan cloud network akan diimplementasikan pada pengujian proses ekstraksi data selanjutnya.

#### IV.3.2 Pengujian Jumlah Item Berdasarkan Lama Waktu Proses

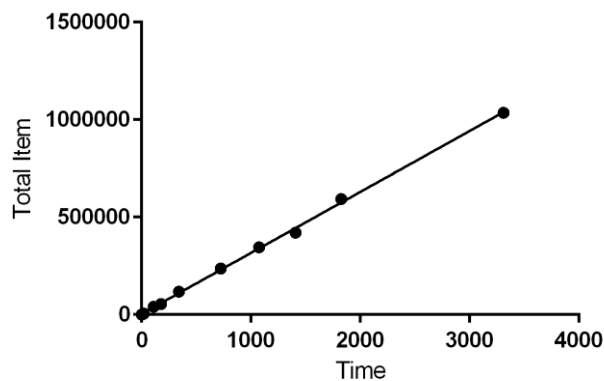
Pengujian ini juga bertujuan untuk mendapatkan perbandingan lama waktu yang dibutuhkan untuk menyelesaikan satu skenario kondisi yang akan diujikan. Pengujian ini dilakukan dengan cloud network seperti yang sudah dijelaskan pada bagian 3.1. Kondisi yang diuji adalah jumlah data yang dimintakan ke API Reddit. Pengujian ini menggunakan perintah yang berbeda untuk setiap command yang diberikan. Variasi yang digunakan dalam pengujian ini adalah jumlah *lap* dan jumlah *batch* scraping. Perintah terminal yang digunakan adalah sebagai berikut (dengan nilai x dan y yang bervariasi).

```
python src/subreddit_downloader.py Hololive --batch-size x --laps y --reddit-id DnpX9tZ075idVbdUuDUGdg --reddit-secret 8-8kRY17k-I0KsHRh7ey-ZCQjks8vA --reddit-username iammedesu --utc-before 1676946171
```

Internet Provider	Proxy/VPN	Ukuran batch (x)	Jumlah lap (y)	Waktu rata-rata setiap lap (menit)	Waktu total (menit)	Item yang didapat
Cloud network	-	10	3	0,03	0,1	157
Cloud network	-	1000	1	18,3	18,3	5957
Cloud network	-	1000	6	18,23	109,4	40486
Cloud network	-	1000	10	17,56	175,6	54697
Cloud network	-	1000	20	17,03	340,6	117885
Cloud network	-	1000	40	18,07	722,8	236340
Cloud network	-	1000	60	17,92	1075,2	345979

Internet Provider	Proxy/VPN	Ukuran batch (x)	Jumlah lap (y)	Waktu rata-rata setiap lap (menit)	Waktu total (menit)	Item yang didapat
Cloud network	-	1000	80	17,6	1407,4	421088
Cloud network	-	1000	100	18,24	1824,7	592810
Cloud network	-	1000	177	18,7	3310,8	1035939

Dari Tabel X, dapat diplot suatu grafik jumlah item yang didapatkan berdasarkan lama waktu satu proses ekstraksi tersebut berlangsung. Didapatkan pula garis regresi linier dari pemodelan grafik tersebut.



Gambar X merupakan grafik plot total item yang didapatkan dari Reddit berdasarkan waktu (menit) yang dihabiskan. Nilai persamaan regresi yang didapatkan adalah:

$$f(x) = 312.8x + 4093$$

Dengan  $f(x)$  adalah jumlah item yang akan didapatkan dan  $x$  adalah waktu yang ditetapkan.

#### IV.3.3 Pengujian Ukuran Data Berdasarkan Lama Waktu Proses

Pengujian ini juga bertujuan untuk mendapatkan perbandingan lama waktu yang dibutuhkan untuk menyelesaikan satu skenario kondisi yang akan diujikan.

Ukuran batch (x)	Jumlah lap (y)	Waktu total (menit)	Jumlah item	Ukuran JSON (MB)	Ukuran CSV (MB)
1000	35				
1000	40	717,6	236340	551,2	58,5

#### IV.4 Persiapan Data

Untuk menunjang proses analisis data, terdapat suatu prosedur untuk memastikan kebenaran, konsistensi, dan kegunaan suatu data yang ada dalam dataset. Tugas Akhir ini menggunakan subreddit “Hololive” sebagai subyek pengujian.

##### IV.4.1 Pembersihan Data

Pembersihan data adalah proses memperbaiki atau menghapus data yang salah, rusak, salah format, duplikat, atau tidak lengkap dalam kumpulan data. Terdapat beberapa data yang kosong dikarenakan kegagalan maupun error dalam proses ekstraksi data menggunakan PRAW. Untuk mengatasinya, seluruh data yang kosong atau NA akan dihapus dari dataset. Digunakan kode program sebagai berikut.

```
df = df.dropna(axis=0, how="any", subset=None, inplace=False)
```

Fungsi kode di atas adalah menghapus baris/raw yang di dalamnya terdapat nilai NA untuk di atribut manapun. Penghapusan ini dimaksudkan untuk membersihkan dataset yang terdapat nilai kosong sehingga tidak mengganggu proses visualisasi data. Kolom yang kosong diasumsikan terjadinya anomali maupun error ketika proses ekstraksi berlangsung.

#### **IV.4.2 Manajemen Stop Word**

Blabla

```

nltk.download('punkt')
nltk.download('stopwords')

STOP_WORDS = stopwords.words()

# removing the emojis
# https://www.kaggle.com/alankritamishra/covid-19-tweet-sentiment-
analysis#Sentiment-analysis
EMOJI_PATTERN = re.compile("[
                                u"\U0001F600-\U0001F64F" # emoticons
                                u"\U0001F300-\U0001F5FF" # symbols &
                                pictographs
                                u"\U0001F680-\U0001F6FF" # transport &
                                map symbols
                                u"\U0001F1E0-\U0001F1FF" # flags (iOS)
                                u"\U00002702-\U000027B0"
                                u"\U000024C2-\U0001F251"
                                "]" + flags=re.UNICODE)

def cleaning(text):
    """
    Convert to lowercase.
    Remove URL links, special characters and punctuation.
    Tokenize and remove stop words.
    """
    text = text.lower()
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('["\'"]+', '', text)

    text = EMOJI_PATTERN.sub(r'', text)

    # removing the stop-words
    text_tokens = word_tokenize(text)
    tokens_without_sw = [word for word in text_tokens if not word in
STOP_WORDS]
    filtered_sentence = (" ").join(tokens_without_sw)
    text = filtered_sentence

    return text

```

## IV.5 Visualisasi Data

Membuat apa yang akan divisualisasi dulu (**desain + 1 prototipe**)

## **BAB V**

### **KESIMPULAN DAN SARAN**

Bagian ini akan menjelaskan kesimpulan dan saran dari hasil pelaksanaan dan pengerjaan tugas akhir. Kesimpulan ditulis untuk menjelaskan dan menjawab beberapa pertanyaan pada rumusan masalah di Bab I. Saran ditulis untuk memberikan insight kepada penelitian dan pengembangan lebih lanjut.

#### **V.1 Kesimpulan**

Adapun kesimpulan dari rumusan masalah tugas akhir ini adalah.

1. Model *web scraper* dapat diterapkan untuk mengambil dan mengolah data yang dapat dimungkinkan untuk menyusun CTI proaktif
2. Informasi yang dapat diperoleh dari proses web scrapping adalah sebagai berikut: naama/identitas aktor, tren keamanan pada linimasa tertentu, dan fokus utama pembicaraan pada rentang waktu tersebut.
3. Model web scraper dapat bekerja sesuai dengan analisis kebutuhan yang sudah dituliskan di Bab III.
4. Visualisasi dapat diterapkan dengan catatan: sumber data dimungkinkan untuk lebih luas dan real-time.

#### **V.2 Saran**

Berikut merupakan saran yang dapat digunakan sebagai panduan pengembangan lebih lanjut.

1. Pengimplementasian machine learning dalam pemberian prediksi tren keamanan pada suatu organisasi dapat diterapkan pada tahap pengolahan data dan visualisasi.
2. Gap antara industri dan akademik dapat diperkecil dengan membuat sebuah purwarupa sistem serupa yang dapat digunakan oleh pelaku organisasi.





## DAFTAR REFERENSI

- . (2016). *A framework for cyber threat hunting*. [online] Available: <https://www.threathunting.net/files/framework-for-threat-hunting-whitepaper.pdf>
- Almohannadi, H., Awan, I., Al Hamar, J., Cullen, A., Disso, J., & Armitage, L. (2018). *Cyber Threat Intelligence from Honeypot Data Using Elasticsearch*. 2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA). pp. 900-906, doi: 10.1109/AINA.2018.00132.
- Amaro, L. J. B., Azevedo, B. W. P., de Mendonca, F. L. L., Giozza, W. F., Albuquerque, R. de, & Villalba, L. J. G. (2022). *Methodological framework to collect, process, analyze and visualize cyber threat intelligence data*. Applied Sciences, 12(3), 1205. <https://doi.org/10.3390/app12031205>.
- Ammari, T., Schoenebeck, S., & Romero, D. (2019). *Self-declared throwaway accounts on Reddit: How platform affordances and shared norms enable parenting disclosure and support*. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–30.
- Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., & Gilbert, E. (2018). *The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales*. Proceedings of the ACM on Human-Computer Interaction, 2, 32.
- Chismon, D., & Ruks, M. (2015). *Threat intelligence: Collecting, analysing, evaluating*. MWR InfoSecurity Ltd, 3(2), 36-42
- Ciobanu, C., Dandurand, L., Grobauer, M., Kacha, B., Kaplan, P., Kompanek, A., & Van Horenbeeck, M. (2014). *Actionable Information for Security Incident Response*. ENISA, Heraklion, Greece
- Farnham, G., & Leune, K. (2013). *Tools and standards for cyber threat intelligence projects*. SANS Institute, 3(2), 25-31
- Feng, B. (2021, August 20). *Threat intelligence sharing: What kind of intelligence to share?* Retrieved October 5, 2022, from <https://www.concordia-h2020.eu/blog-post/threat-intelligence-sharing/>
- Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). *Web data extraction, applications and techniques: A survey*. Knowledge-Based Systems. 70. 301–323. <https://doi.org/10.1016/j.knosys.2014.07.007>
- Fiesler, C., Beard, N., & Keegan, B. C. (2020). *No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service*. Proceedings of the International AAAI Conference on Web and Social Media, 14, 187–196.

- Fu, T., Abbasi, A., & Chen, H. (2010). *A Focused Crawler for Dark Web Forums*. Journal of the American Society for Information Science and Technology. [online] Available: <https://doi.org/10.1002/asi>
- Hilbert, M. (2015). *Big data for development: A review of promises and challenges*. Development Policy Review. pp. 10-07.
- Jiang, J., Song, X., Yu, N., & Lin, C. Y. (2014). *FoCUS?: Learning to Crawl Web Forums*. IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 6, pp. 1293-1306
- Johnson, C., Badger, L., Waltermire, D., Snyder, J., & Skorupka, C. (2016, October). *Guide to cyber threat information sharing - NIST*. National Institute of Standards and Technology. Retrieved October 1, 2022, from <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-150.pdf>
- Lee, R., Assante, M., & Conway, T. (2016, March). *Analysis of the Cyber Attack on the Ukrainian Power Grid*. Electricity Information Sharing and Analysis Center. Retrieved October 5, 2022, from [https://paper.seebug.org/papers/APT/APT\\_CyberCriminal\\_Campagin/2016/2016.03.18.Analysis\\_of\\_the\\_Cyber\\_Attack\\_on\\_the\\_Ukrainian\\_Power\\_Grid/E-ISAC\\_SANS\\_Ukraine\\_DUC\\_5.pdf](https://paper.seebug.org/papers/APT/APT_CyberCriminal_Campagin/2016/2016.03.18.Analysis_of_the_Cyber_Attack_on_the_Ukrainian_Power_Grid/E-ISAC_SANS_Ukraine_DUC_5.pdf)
- Marres, N., & Weltevrede, E. (2013). *Scraping the social?* Journal of Cultural Economy. 6(3). 313–335. <https://doi.org/10.1080/17530350.2013.772070>
- Medvedev, A. N., Lambiotte, R., & Delvenne, J.-C. (2019). *The Anatomy of Reddit: An Overview of Academic Research*. Springer Proceedings in Complexity, 183–204. doi:10.1007/978-3-030-14683-2\_9
- Nunes, E. (2016). *Darknet and deepnet mining for proactive cybersecurity threat intelligence*. 2016 IEEE Conference on Intelligence and Security Informatics (ISI). pp. 7-12, doi: 10.1109/ISI.2016.7745435.
- Parvez, M. S., Tasneem, K. S., Rajendra, S. S., & Bodke, K. R. (2018). *Analysis of different web data extraction techniques*. 2018 International Conference on Smart City and Emerging Technology (ICSCET). <https://doi.org/10.1109/icscet.2018.8537333>
- Pavkovic, M. & Protic, J. (2013). *Intelligent crawler for web forums based on improved regular expressions*. 2013 21st Telecommunications Forum Telfor TELFOR 2013 - Proceedings of Papers, pp. 817-820, [online] Available: <https://doi.org/10.1109/TELFOR.2013.6716355>.
- Portokalidis, G., Slowinska, A., & Bos, H. (2006). *Argos: an emulator for fingerprinting zero-day attacks for advertised honeypots with automatic signature generation*. ACM SIGOPS Operating Systems Review, ACM. vol. 40. pp. 15-27

- Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). *Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics*. *Social Media + Society*, 7(2). <https://doi.org/10.1177/20563051211019004>
- Reddit.com. (2020). *User agreement—October 15, 2020—Reddit*. Reddit User Agreement. <https://www.redditinc.com/policies/user-agreement-october-15-2020>
- Reddit.com. (2021, January 17). *Advertising—Audience—Reddit*. Discover what makes Reddit ads unique. <https://web.archive.org/web/20210117184818/https://www.redditinc.com/advertising/audience>
- Samtani, S., Chinn, R., Chen, H., & Nunamaker, J. F. (2017). *Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence*. *Journal of Management Information Systems*, 34(4), 1023–1053. doi:10.1080/07421222.2017.1394049
- Shackleford, D. (2015). *Who's using cyberthreat intelligence and how?* SANS Institute. From [www.sans.org/reading-room/whitepapers/analyst/cyberthreat-intelligence-how35767](http://www.sans.org/reading-room/whitepapers/analyst/cyberthreat-intelligence-how35767)
- Song, J., Takakura, H., Okabe, Y., Eto, M., Inoue, D., & Nakao, K. (2011). *Statistical analysis of honeypot data and building of kyoto 2006+ dataset for nids evaluation*. *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security* pp. 29-36
- Wagner, T. D., Mahbub, K., Palomar, E., & Abdallah, A. E. (2019). *Cyber threat intelligence sharing: Survey and research directions*. *Computers & Security*, 87. <https://doi.org/10.1016/j.cose.2019.101589>
- Williams, R., Samtani, S., Patton, M., and Chen, H. (2018). *Incremental Hacker Forum Exploit Collection and Classification for Proactive Cyber Threat Intelligence: An Exploratory Study*. 2018 IEEE International Conference on Intelligence and Security Informatics (ISI). pp. 94-99. doi: 10.1109/ISI.2018.8587336

## **Lampiran A. Contoh Judul Lampiran**

### **A.1 Contoh Judul Anak Lampiran**

Contoh anak lampiran