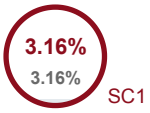
Similarity Report

Metadata

Name of the organization
Universitas Muhammadiyah Sidoarjo
Title
211080200045_Naufal Ariq Wijaya_Artikel HAKI
Author(s) Coordinator
perpustakaan umsidaprist
Organizational unit
Perpustakaan

Record of similarities

SCs indicate the percentage of the number of words found in other texts compared to the total number of words in the analysed document. Please note that high coefficient values do not automatically mean plagiarism. The report must be analyzed by an authorized person.



25	6553	50600
The phrase length for the SC 2	Length in words	Length in characters

Alerts

In this section, you can find information regarding text modifications that may aim at temper with the analysis results. Invisible to the person evaluating the content of the document on a printout or in a file, they influence the phrases compared during text analysis (by causing intended misspellings) to conceal borrowings as well as to falsify values in the Similarity Report. It should be assessed whether the modifications are intentional or not.

Characters from another alphabet		0
Spreads	A→	0
Micro spaces		1
Hidden characters		0
Paraphrases (SmartMarks)	a	12

Active lists of similarities

This list of sources below contains sources from various databases. The color of the text indicates in which source it was found. These sources and Similarity Coefficient values do not reflect direct plagiarism. It is necessary to open each source, analyze the content and correctness of the source crediting.

The 10 longest fragments — Color of the text

NO	TITLE OR SOURCE URL (DATABASE)	NUMBER OF IDENTICAL WORDS (FRAGMENTS)
1	https://stackoverflow.com/questions/77895157/how-can-i-solve-userwarning-set-ticklabels-should-only-be-used-with-a-fixed	27 0.41 %
2	https://123dok.com/document/q05o797v-recognition-indonesia-berbasis-classification-rizkiyanto-matematika-pengetahuan-universitas.html	21 0.32 %
3	Efektivitas Penerapan Konsep Keamanan Digital oleh Mahasiswa Rita Gani,Sarry Shafina Saraswati;	18 0.27 %

4	https://cyberleninka.ru/article/n/natural-language-processing-and-fiction-text-basis-for-corpus-research	18 0.27 %
5	https://cyberleninka.ru/article/n/natural-language-processing-and-fiction-text-basis-for-corpus-research	16 0.24 %
6	Розробка тестового середовища для моделювання мережевих атак [redacted] 6/17/2024 National University "Zaporizhzhia Polytechnic" (Кафедра "Інформаційна безпека та наноелектроніка")	15 0.23 %
7	https://www.ajol.info/index.php/sej/article/download/274626/259255/645838	12 0.18 %
8	https://www.ajol.info/index.php/sej/article/download/274626/259255/645838	11 0.17 %
9	Evaluation of the Cochrane Consumers and Communication Group's systematic review priority-setting project Allison Tong, Sophie Hill,Anneliese Synnot, Rebecca Ryan;	11 0.17 %
10	Rajamäe_Soosaar_ITmitteinformaatikutele_2024.pdf [redacted] 5/30/2024 Tartu Ülikool (University of Tartu)	9 0.14 %

from RefBooks database (0.52 %)

NO	TITLE	NUMBER OF IDENTICAL WORDS (FRAGMENTS)
Source: Paperity		
1	Efektivitas Penerapan Konsep Keamanan Digital oleh Mahasiswa Rita Gani,Sarry Shafina Saraswati;	18 (1) 0.27 %
2	Evaluation of the Cochrane Consumers and Communication Group's systematic review priority-setting project Allison Tong, Sophie Hill,Anneliese Synnot, Rebecca Ryan;	11 (1) 0.17 %
3	Application of Named Entity Recognition via Twitter on SpaCy in Indonesian (Case Study : Power Failure in the Special Region of Yogyakarta) Suadaa Lya Hulliyyatus, Ibnu Santoso,Yanti Rizka Maulida;	5 (1) 0.08 %

from the home database (0.00 %)

NO	TITLE	NUMBER OF IDENTICAL WORDS (FRAGMENTS)

from the Database Exchange Program (0.37 %)

NO	TITLE	NUMBER OF IDENTICAL WORDS (FRAGMENTS)
1	Розробка тестового середовища для моделювання мережевих атак [redacted] 6/17/2024 National University "Zaporizhzhia Polytechnic" (Кафедра "Інформаційна безпека та наноелектроніка")	15 (1) 0.23 %
2	Rajamäe_Soosaar_ITmitteinformaatikutele_2024.pdf [redacted] 5/30/2024 Tartu Ülikool (University of Tartu)	9 (1) 0.14 %

from the Internet (2.27 %)

NO	SOURCE URL	NUMBER OF IDENTICAL WORDS (FRAGMENTS)
1	https://cyberleninka.ru/article/n/natural-language-processing-and-fiction-text-basis-for-corpus-research	34 (2) 0.52 %
2	**https://stackoverflow.com/questions/77895157/how-can-i-solve-userwarning-set-ticklabels-should-only-be-used-with-a-fixed**	**27 (1) 0.41 %**
3	https://archive.umsida.ac.id/index.php/archive/preprint/download/5087/36261/40797	24 (4) 0.37 %

4	https://www.ajol.info/index.php/sej/article/download/274626/259255/645838	23 (2) 0.35 %
5	https://123dok.com/document/q05o797v-recognition-indonesia-berbasis-classification-rizkiyanto-matematika-pengetahuan-universitas.html	21 (1) 0.32 %
6	https://sites.google.com/view/parameshacharibd/publications	14 (2) 0.21 %
7	https://link.springer.com/chapter/10.1007/978-981-99-5994-5_17	6 (1) 0.09 %

List of accepted fragments (no accepted fragments)

NO	CONTENTS	NUMBER OF IDENTICAL WORDS (FRAGMENTS)

Leveraging OSINT for Identifying and Mitigating Cybercrime Threats in Online Media Environments
[Pemanfaatan Osint Untuk Identifikasi Dan Mitigasi Ancaman Kejahatan Cyber Di Lingkungan Media Daring]

Naufal Ariq Wijaya1), Arif Senja Fitrani2), Hindarto3), M. **Alfan** Rosid 4),
1) Program Studi **Informatika, Universitas Muhammadiyah** Sidoarjo, Indonesia
2) Program **Studi Informatika, Universitas Muhammadiyah Sidoarjo,** Indonesia
3) Program **Studi Informatika, Universitas Muhammadiyah Sidoarjo,** Indonesia
4) Program **Studi Informatika, Universitas Muhammadiyah** Sidoarjo, Indonesia

Page | 1
2 | Page
Page | 3

Abstract. Cybercrime in online media, such as the misuse of data from open sources, poses a serious threat in the digital era. This study utilizes an Open-Source Intelligence (OSINT) approach to identify and mitigate such threats by developing an automated tool designed to collect and analyze public data from various search engines, including Google, DuckDuckGo, Yahoo, AOL, Twitter, Telegram, and Bing News. The tool is developed using the Python programming language due to its flexibility and the strong ecosystem of libraries for data analysis and text processing. Data extraction combines Named Entity Recognition (NER) powered by spaCy and regular expressions (regex) to detect and classify key metadata such as personal names, organizations, email addresses, phone numbers, social media accounts, and website domains. The analysis results are visualized through bar charts, line graphs, and network diagrams, and presented in an interactive HTML report using Bootstrap 5, making it easy for users to interpret and explore the findings. All data **used in this study is publicly accessible, and** the research strictly adheres to ethical standards in cyber investigations. The findings demonstrate that an automated OSINT approach using Python is effective in identifying exposed public information and supports mitigation strategies through structured entity analysis. This research contributes to the development of practical and academic OSINT tools for enhancing cybersecurity practices in digital environments.

Keywords - OSINT, Cybercrime, Online Media, Threat Identification, Threat Mitigation, Metadata Extraction, Data Visualization, Cybersecurity

Abstrak. Kejahatan siber di media daring, seperti penyalahgunaan data dari sumber terbuka, menjadi ancaman serius di era digital. Penelitian ini memanfaatkan pendekatan Open-Source Intelligence (OSINT) untuk mengidentifikasi dan memitigasi ancaman tersebut melalui pengembangan tool otomatis yang dirancang untuk mengumpulkan serta menganalisis data publik dari berbagai mesin pencari, termasuk Google, DuckDuckGo, Yahoo, AOL, Twitter, Telegram, dan Bing News. Tool ini dibangun menggunakan bahasa pemrograman Python karena sifatnya yang fleksibel dan dukungan ekosistem pustaka (library) yang kuat untuk kebutuhan analisis data dan pemrosesan teks. Proses ekstraksi data dilakukan dengan menggabungkan teknik Named Entity Recognition (NER) berbasis spaCy dan regular expression (regex) untuk mendeteksi dan mengklasifikasi metadata penting seperti nama individu, organisasi, alamat email, nomor telepon, akun media sosial, dan domain situs web. Seluruh hasil dianalisis dan divisualisasikan melalui berbagai grafik, seperti diagram batang, garis, dan network graph, serta ditampilkan dalam laporan HTML interaktif berbasis Bootstrap 5, sehingga memudahkan pengguna dalam membaca serta menavigasi hasil analisis. Tool ini hanya memanfaatkan data publik (tanpa akses ilegal), serta menjunjung tinggi prinsip-prinsip etika penelitian siber. Temuan menunjukkan bahwa pendekatan OSINT yang diotomatisasi menggunakan Python dapat secara efektif mengidentifikasi informasi yang terekspos di ruang publik dan mendukung langkah-langkah mitigasi melalui analisis entitas yang terstruktur. Penelitian ini memberikan kontribusi pada pengembangan alat bantu OSINT untuk keamanan siber, baik di ranah akademik maupun praktik profesional digital.

Kata Kunci - OSINT, Kejahatan Siber, Media Daring, Identifikasi Ancaman, Mitigasi Ancaman, Ekstraksi Metadata, Visualisasi Data, Keamanan Siber

1. I. Pendahuluan

Pesatnya perkembangan teknologi telah memicu lonjakan kejahatan siber di media daring, yang mengancam individu dan organisasi [1]. Kejahatan siber yang semakin canggih menuntut strategi efektif untuk mengidentifikasi dan memitigasi ancaman. Penelitian ini memanfaatkan Open-Source Intelligence (OSINT), sebuah metode pengumpulan data publik yang sesuai dengan hukum [2], dikombinasikan dengan teknik footprinting untuk mengatasi tantangan tersebut.

Footprinting merujuk pada proses pengumpulan informasi secara sistematis tentang target, baik individu maupun organisasi, untuk menganalisis jejak digital mereka [3]. Penelitian ini bertujuan untuk mengidentifikasi pola aktivitas entah itu pelaku kejahatan siber atau jejak digital seseorang yang ada di

internet, sehingga mendukung pengembangan strategi mitigasi ancaman yang efektif. Untuk mencapai tujuan ini, penelitian ini mengembangkan tool berbasis Python yang mengintegrasikan berbagai sumber data publik, termasuk mesin pencari seperti Google, DuckDuckGo, Yahoo, dan AOL, serta platform seperti Twitter, Telegram, dan Bing News.

Menggunakan spaCy, sebuah pustaka untuk Named Entity Recognition (NER) [4], yang menggabungkan OSINT dengan pembelajaran mesin untuk mengekstrak metadata penting, seperti nama individu, organisasi, alamat email, nomor telepon, akun media sosial, dan domain situs web. Pendekatan ini memperkuat kemampuan deteksi dan analisis ancaman kejahatan siber di lingkungan media daring, memberikan kontribusi signifikan bagi penelitian akademik dan aplikasi keamanan siber praktis [5].

Tools ini menghasilkan output laporan HTML interaktif berbasis Bootstrap 5 yang responsif dan user-friendly [6], dilengkapi dengan visualisasi data untuk mendukung analis dan investigator dalam memahami ancaman siber. Fitur utama meliputi grafik metadata yang menampilkan frekuensi kemunculan entitas, seperti nama individu atau organisasi, dengan tombol "Lihat Detail Entitas" untuk menunjukkan entitas dominan berdasarkan NER dan "Lihat Semua Metadata" untuk menampilkan seluruh metadata yang diekstrak, termasuk email, nomor telepon, dan akun media sosial.

Selain itu, grafik domain menggambarkan frekuensi kemunculan domain dari hasil pencarian, membantu mengidentifikasi sumber data utama. Grafik kata umum menyoroti sepuluh kata teratas dalam deskripsi pencarian, mengungkap pola konten yang relevan. Diagram karakter dan digit menyajikan total karakter dan digit dalam metadata untuk analisis teknis, sementara grafik email dan nomor telepon menampilkan jumlah kontak yang ditemukan. Visualisasi network graph mengilustrasikan hubungan antara hasil pencarian dan entitas, seperti keterkaitan antara domain dan nama individu, mempermudah analisis jejaring. Terakhir, tool ini mendukung pemeriksaan manual email melalui HaveIBeenPwned untuk mendeteksi potensi kebocoran data [7]. Tools ini dirancang untuk mematuhi Undang-Undang Perlindungan Data Pribadi (UU PDP) dengan hanya memproses data publik dan menjunjung etika penelitian [8], memastikan penggunaan yang bertanggung jawab dan analisis ancaman yang terstruktur melalui laporan HTML yang interaktif.

2. II. Metode

1. Pendekatan Penelitian

Penelitian ini menerapkan pendekatan Open-Source Intelligence (OSINT) untuk mengumpulkan data publik dari berbagai sumber di internet, yang dikombinasikan dengan teknik footprinting guna menganalisis jejak digital target, baik individu maupun organisasi. OSINT memungkinkan pengumpulan informasi yang tersedia secara legal, seperti hasil pencarian mesin pencari atau postingan media sosial, untuk mendeteksi pola sebuah entitas [9]. Footprinting, sebagai metode pengumpulan informasi sistematis, digunakan untuk memetakan jejak digital seperti alamat email, nomor telepon, atau domain situs web yang terkait dengan aktivitas siber [10]. Pendekatan ini mengintegrasikan Named Entity Recognition (NER) berbasis pembelajaran mesin atau machine learning untuk mengidentifikasi entitas seperti nama atau organisasi, serta ekstraksi pola menggunakan regular expression (regex) untuk mengekstrak data terstruktur seperti no telpon, email dan link media sosial. Kombinasi teknik ini memastikan deteksi data apa saja yang dapat dikumpukan dan dianalisa lebih mendalam.

2. Pengembangan Tool

Pengembangan tool dalam penelitian ini bertujuan untuk mencari sebuah jejak digital melalui pendekatan Open-Source Intelligence (OSINT). Tool ini dibangun menggunakan bahasa pemrograman Python, yang dipilih karena fleksibilitas dan dukungan ekosistem library-nya untuk analisis data dan pembelajaran mesin [11]. Fungsi utama tool meliputi pengumpulan data publik, ekstraksi metadata, analisis jejak digital, dan penyajian visualisasi dalam laporan HTML interaktif, memungkinkan analis dan investigator untuk memahami jejak digital secara efektif.

Arsitektur tool terdiri dari dua modul utama yang saling terintegrasi. Modul pertama, yang diimplementasikan melalui skrip search_engine.py, bertugas mengumpulkan data dari sumber publik seperti mesin pencari (Google, DuckDuckGo, Yahoo, AOL) dan platform (Twitter, Telegram, Bing News) berdasarkan kueri relevan, misalnya “kejahatan siber Indonesia” atau “NAMA SESEORANG”. Modul kedua, diimplementasikan melalui skrip main.py, mengelola pemrosesan data, ekstraksi metadata, dan pembuatan laporan visual. Diagram arsitektur tool, yang menggambarkan alur dari pengumpulan data hingga pelaporan, disajikan pada Gambar 2.1.

Gambar 2. SEQ Gambar * ARABIC 1 Flowchart Arsitektur Tool OSINT.

Fungsionalitas tool didukung oleh beberapa library Python. Library requests dan BeautifulSoup memungkinkan akses dan parsing halaman web untuk pengumpulan data yang akurat. Library pandas digunakan untuk analisis, sedangkan matplotlib dan seaborn menghasilkan visualisasi seperti grafik frekuensi domain. Library Jinja2 mendukung pembuatan laporan HTML dinamis berbasis Bootstrap 5, yang menawarkan antarmuka responsif lintas perangkat [12]. Proses ekstraksi dan analisis pola dijelaskan pada subbab berikut untuk menjaga kejelasan alur kerja.

3. Pengumpulan Data

Pada tahap ini, proses pengumpulan data dilakukan menggunakan metode Open-Source Intelligence (OSINT) yang diotomatisasi melalui sebuah tool yang dikembangkan dalam penelitian ini. Tool tersebut bekerja berdasarkan teknik footprinting, yaitu metode untuk melacak dan mengidentifikasi jejak digital yang mungkin terkait dengan individu, organisasi, atau peristiwa tertentu.

Sebagai ilustrasi, tool dapat digunakan untuk mencari berita lama atau informasi mengenai seseorang, yang kemungkinan memiliki keterkaitan digital baik berupa nama, akun, maupun aktivitas di internet. Pencarian dilakukan secara otomatis terhadap berbagai sumber terbuka, seperti mesin pencari (Google, DuckDuckGo, dan lainnya), media sosial, hingga portal berita. Dari hasil tersebut, informasi penting seperti judul, deskripsi, dan metadata laman akan dikumpulkan.

Seluruh aktivitas dilakukan menggunakan tool memastikan efisiensi serta cakupan data yang luas. Informasi yang telah terkumpul kemudian akan diproses lebih lanjut pada tahap ekstraksi metadata. Proses ini penting untuk menyaring dan menyiapkan data agar dapat dianalisis secara lebih mendalam pada tahap-tahap berikutnya.ada tahap ini, tool akan melakukan pencarian informasi dari berbagai sumber mesin pencari seperti Google, DuckDuckGo, Yahoo, dan AOL, serta dari media sosial seperti Twitter dan portal berita Bing News [13]. Tool secara otomatis mencari data yang relevan berdasarkan kata kunci yang diinginkan, mencakup elemen-elemen penting seperti judul, deskripsi, dan metadata dari laman website yang memuat kata kunci tersebut. Setelah pengumpulan data selesai, proses selanjutnya adalah ekstraksi metadata, yaitu pengambilan informasi kunci dari data yang

terkumpul untuk mendukung analisis lebih lanjut.

4. Ekstraksi Metadata
Ekstraksi metadata bertujuan mengidentifikasi informasi kunci dari data terkumpul untuk mendeteksi pola entitas. Library spaCy dengan model en_core_web_trf digunakan untuk Named Entity Recognition (NER), mengenali entitas seperti nama individu (contoh: Sundar Pichai), organisasi (contoh: Google), dan lokasi (contoh: New York) dari teks pencarian [14]. Contohnya, dari hasil pencarian “Siapa CEO Google Sekarang”, spaCy mengidentifikasi “Google” sebagai organisasi, “Sundar Pichai” sebagai individu, dan “New York” sebagai lokasi, memungkinkan pelacakan jejak digital secara akurat.

Selain itu, regular expression (regex) mengekstrak pola spesifik seperti alamat email (contoh: s**********i@google.com), nomor telepon (contoh: +1 234-6578-91011), dan akun media sosial (contoh: @sundarpichai). Regex menangani variasi teks dari deskripsi atau posting media sosial, menghasilkan data terstruktur dalam format CSV atau JSON. Validasi melalui pengujian kueri memastikan keandalan ekstraksi, dengan optimasi untuk format data yang beragam. Contoh diagram cara kerja ekstraksi metadata ditampilkan pada Gambar 2.2.

Gambar 2. SEQ Gambar * ARABIC 2 Diagram ekstraksi metadata menggunakan Named Entity Recognition (NER).

5. Analisis dan Visualisasi Data
Analisis metadata memiliki peran krusial dalam mengidentifikasi pola jejak digital serta mengurai berbagai jenis data yang tersedia di internet. Data yang berhasil diekstraksi kemudian diproses menggunakan library pandas untuk menghitung frekuensi kemunculan entitas, domain, dan kata kunci, sehingga menghasilkan gambaran kuantitatif terkait distribusi dan intensitas elemen-elemen tersebut.

Hasil analisis tersebut disajikan dalam bentuk laporan HTML interaktif yang dikembangkan dengan framework Bootstrap 5, memastikan antarmuka yang responsif dan mudah diakses melalui berbagai perangkat, sebagaimana dijelaskan dalam subbab 2.2 [15]. Laporan ini menyediakan berbagai fitur visualisasi untuk memudahkan interpretasi data. Grafik frekuensi entitas dilengkapi dengan tombol "Lihat Detail Entitas" yang memungkinkan eksplorasi lebih mendalam, serta tombol "Lihat Semua Metadata" untuk menampilkan informasi tambahan seperti alamat email dan akun media sosial.

Selain itu, grafik domain menampilkan distribusi situs web yang terdeteksi, sementara grafik kata umum mengungkap sepuluh kata paling dominan dalam kumpulan data. Diagram karakter dan digit memberikan analisis statistik teknis terkait komposisi metadata, sedangkan grafik kontak menyajikan jumlah email dan nomor telepon yang berhasil diidentifikasi. Untuk memperkaya analisis hubungan antar-entitas, laporan juga menyertakan network graph yang memvisualisasikan koneksi antara berbagai elemen, seperti hubungan antara situs web dan individu tertentu. Pendekatan ini mempermudah pemahaman terhadap pola jejaring, khususnya dalam konteks evaluasi terhadap langkah yang harus di ambil kedepannya.

6. Pemeriksaan Kebocoran Data
Pemeriksaan kebocoran data merupakan bagian penting dalam proses identifikasi risiko keamanan informasi, khususnya pada tahap analisis metadata. Selama proses ini, tool yang dikembangkan dapat mendeteksi informasi sensitif seperti alamat email dari berbagai sumber publik, misalnya dari situs web atau unggahan media sosial.

Alamat email yang terdeteksi kemudian diperiksa secara manual menggunakan layanan Have I Been Pwned untuk mengetahui apakah data tersebut pernah terlibat dalam insiden kebocoran data [16]. Pemeriksaan ini memberikan gambaran awal mengenai potensi kerentanan keamanan yang mungkin melekat pada informasi yang ditemukan.

Untuk menjaga integritas proses dan mencegah penyalahgunaan, data yang dikumpulkan disimpan sementara dalam format file HTML dan hanya dapat di akses pada lokal file. Selain itu, pengujian rutin juga dilakukan untuk memastikan akurasi pendeteksian serta mendukung pendekatan investigasi yang lebih proaktif dan terarah.

3. III. Hasil dan Pembahasan
1. Tampilan Menu Utama
Gambar 3.1 menunjukkan tampilan awal saat tool dijalankan menggunakan platform Google Colaboratory (Colab). Tool ini diintegrasikan langsung dengan Google Drive untuk mempermudah akses file dan direktori proyek, khususnya folder kerja yang berisi skrip utama dan kebutuhan pustaka pendukung.

Langkah pertama dimulai dengan mounting Google Drive ke lingkungan kerja Colab menggunakan perintah drive.mount('/content/drive'). Ini memungkinkan kita mengakses folder proyek yang sebelumnya telah disimpan, dalam hal ini folder bernama OSINT_All. Setelah itu, dilakukan perpindahan direktori kerja (%cd) ke dalam folder tersebut agar semua perintah dan pemanggilan file berjalan dari lokasi yang sesuai.

Sebelum menjalankan skrip utama (main.py), beberapa pustaka yang dianggap tidak kompatibel atau bisa menyebabkan konflik versi terlebih dahulu dihapus menggunakan pip uninstall. Kemudian dilakukan instalasi ulang semua dependensi melalui file requirements.txt, agar lingkungan kerja benar-benar bersih dan sesuai dengan kebutuhan tool.

Terakhir, skrip main.py dijalankan. Skrip ini memuat seluruh logika pemrosesan, mulai dari pengumpulan data OSINT, ekstraksi metadata, hingga proses identifikasi entitas seperti nama, organisasi, atau alamat email.

Tampilan ini secara keseluruhan mencerminkan antarmuka kerja awal yang bersifat otomatis dan terstruktur. Menggunakan Google Colab juga memberi keuntungan praktis, karena pengguna tidak perlu melakukan instalasi di perangkat lokal, cukup menggunakan browser dan koneksi internet..

Gambar 3. SEQ Gambar * ARABIC\r 1 1 Tampilan awal antarmuka Google Colab saat tool OSINT diinisialisasi. Proses mencakup mounting Google Drive, perpindahan direktori ke folder proyek, instalasi dependensi, dan eksekusi skrip utama.

2. Tampilan Pilihan Pencarian
Memperlihatkan antarmuka input pada tahap awal proses pencarian informasi menggunakan tool OSINT. Pengguna diberikan berbagai opsi mesin pencari yang dapat digunakan secara terpisah maupun digabungkan, seperti Google, DuckDuckGo, Yahoo, hingga AOL. Selain itu, tool ini juga mendukung pencarian melalui media sosial dan portal berita seperti Twitter (Xstalk), TikTok (melalui Urlebird), Telegram, serta Bing News.

Pada contoh ini, pengguna memilih opsi ke-5, yaitu gabungan dari semua mesin pencari, untuk mendapatkan hasil yang lebih luas. Selanjutnya, pengguna memasukkan kueri pencarian berupa "Sundar Pichai AND Email", yang menunjukkan bahwa tool akan mencari hasil yang mengandung kedua kata kunci tersebut. Pengguna juga diberi opsi untuk memfilter pencarian berdasarkan tipe file (seperti PDF atau XLSX) dan domain tertentu (misalnya .go.id atau .ac.id), meskipun dalam contoh ini, kedua filter tersebut dikosongkan agar pencarian dilakukan secara umum untuk mencakup semua pencarian.

Gambar 3. SEQ Gambar * ARABIC 2 Tampilan antarmuka input pencarian pada tool OSINT, di mana pengguna dapat memilih mesin pencari, memasukkan kueri, serta menyaring hasil berdasarkan tipe file dan domain tertentu.

3. Tampilan Ringkasan Metadata Hasil Pencarian
Pada tabel di bawah ini ditampilkan output hasil ekstraksi metadata dari proses pencarian informasi menggunakan tool OSINT. Proses ini memanfaatkan Named Entity Recognition (NER) dengan bantuan library SpaCy untuk mengidentifikasi berbagai entitas penting seperti alamat email, nomor telepon, akun media sosial, nama tokoh, organisasi, hingga lokasi geografis.

Sebagai contoh, sistem berhasil menemukan alamat email s****r@google.com, nomor telepon, serta akun LinkedIn yang diduga milik Sundar Pichai. Semua entitas tersebut secara otomatis dikategorikan ke dalam jenis seperti PERSON, ORG, GPE, DATE, dan lainnya, lengkap dengan jumlah kemunculannya dalam data sumber.

Hasil ekstraksi menunjukkan bahwa kategori ORG (organisasi) paling dominan, dengan entitas “Google” muncul sebanyak 240 kali, diikuti oleh PERSON dengan nama “Sundar Pichai” yang disebutkan 229 kali. Hal ini mengindikasikan bahwa fokus utama dari konten yang dianalisis adalah pada individu dan organisasi tersebut.

Selain itu, terdapat pula daftar domain asal data seperti xstalk.com, msn.com, dan urlebird.com. Frekuensi kemunculan dari tiap domain dihitung untuk memberikan gambaran seberapa banyak informasi terkait subjek diambil dari masing-masing situs web.

Metadata Summary: [{'type': 'Emails', 'entity': 's****r@google.com', 'count': '6 kali'}, {'type': 'Phone Numbers', 'entity': '6**-***-**70', 'count': '1 kali'}, {'type': 'Social Media', 'entity': 'https://www.linkedin.com/in/sundarpichai/.', 'count': '1 kali'}, {'type': 'PERSON', 'entity': 'Sundar Pichai', 'count': '229 kali'}, {'type': 'NORP', 'entity': 'Indian-American', 'count': '2 kali'}, {'type': 'FAC', 'entity': 'the E. Barrett Prettyman Federal Courthouse', 'count': '1 kali'}, {'type': 'ORG', 'entity': 'Google', 'count': '240 kali'}, {'type': 'GPE', 'entity': 'US', 'count': '9 kali'}, {'type': 'LOC', 'entity': 'Tidak ada entitas ditemukan', 'count': '0 kali'}, {'type': 'PRODUCT', 'entity': 'Gemini', 'count': '16 kali'}, {'type': 'EVENT', 'entity': 'Tidak ada entitas ditemukan', 'count': '0 kali'}, {'type': 'WORK_OF_ART', 'entity': 'EMAIL', 'count': '1 kali'}, {'type': 'LAW', 'entity': 'Code of', 'count': '1 kali'}, {'type': 'LANGUAGE', 'entity': 'English', 'count': '1 kali'}, {'type': 'DATE', 'entity': 'Wednesday', 'count': '18 kali'}, {'type': 'TIME', 'entity': 'morning', 'count': '3 kali'}, {'type': 'PERCENT', 'entity': '6%', 'count': '2 kali'}, {'type': 'MONEY', 'entity': '$10.73 million', 'count': '2 kali'}, {'type': 'QUANTITY', 'entity': 'Tidak ada entitas ditemukan', 'count': '0 kali'}, {'type': 'ORDINAL', 'entity': 'first', 'count': '9 kali'}, {'type': 'CARDINAL', 'entity': '12,000', 'count': '8 kali'}] ax.set_xticklabels(categories, rotation=45, ha='right') Domain Counts: Counter({'xstalk.com': 44, 'www.msn.com': 20, 'urlebird.com': 10, 'timesofindia.indiatimes.com': 8, 'www.crn.com': 7, 'www.businessinsider.com': 4, 'www.cnbc.com': 4, 'www.moneycontrol.com': 4, 'www.indiatoday.in': 4, 'www.linkedin.com': 3, 'www.theverge.com': 3, 'en.wikipedia.org': 3, 'arstechnica.com': 3, 'www.bloomberg.com': 3, 'www.techrepublic.com': 3, 'www.quora.com': 2, 'www.justice.gov': 2, 'nypost.com': 2, 'www.axios.com': 2, 'www.nytimes.com': 2, 'www.livemint.com': 2, 'www.elliott.org': 2, 'www.business-standard.com': 2, 'indianexpress.com': 2, 'www.benzinga.com': 2, 'www.businesstoday.in': 2, 'www.afr.com': 2, 'www.entrepreneur.com': 2, 'www.aloye.com': 2, 'www.india.com': 2, 'www.youtube.com': 2, 'www.ceoinfluencers.com': 2, 'www.npr.org': 2, 'www.theinformation.com': 2, 'www.reuters.com': 2, 'economictimes.indiatimes.com': 2, 'thehill.com': 2, 'in.mashable.com': 2, 'www.theglobeandmail.com': 2, 'www.nbcnewyork.com': 2, 'searchengineland.com': 2, 'www.neowin.net': 2, 'blog.google': 1, 'atal.substack.com': 1, 'www.reddit.com': 1, 'www.ceoemail.com': 1, 'support.google.com': 1, 'twitter.com': 1, 'actions.eko.org': 1, 'rocketreach.co': 1, 'medium.com': 1, 'www.instagram.com': 1, 'www.inc.com': 1, 'fortune.com': 1, 'www.hindustantimes.com': 1, 'mashable.com': 1, 'm.economictimes.com': 1, 'www.saleshandy.com': 1, 'abc7news.com': 1, 'gmail.googleblog.com': 1, 'www.bbc.com': 1, 'www.economicliberties.us': 1, 'philanthropynewsdigest.org': 1, 'www.timesnownews.com': 1, 'www.ndtv.com': 1, 'www.outlookbusiness.com': 1, 'dl.acm.org': 1, 'www.fastcompany.com': 1, 'www.androidcentral.com': 1, 'cio.eletsonline.com': 1, 'www.thekurzweillibrary.com': 1, 'www.kuow.org': 1, 'www.techemails.com': 1, 'x.com': 1, 'www.legaldive.com': 1, 'www.hawley.senate.gov': 1, 'safety.google': 1, 'dojmt.gov': 1, 'www.yahoo.com': 1, 'theconversation.com': 1, 'www.vice.com': 1, 'sapiengraph.com': 1, 'www.indiatvnews.com': 1, 'www.cbsnews.com': 1, 'www.mprnews.org': 1, 'www.wheresyoured.at': 1, '9to5google.com': 1, 'people.com': 1, 'abc.xyz': 1, 'gizmodo.com': 1, 'www.dailymail.co.uk': 1, 'phys.org': 1, 'financialpost.com': 1, 'mezha.media': 1, 'www.geekwire.com': 1, 'www.cbs17.com': 1, 'www.the-independent.com': 1, 'venturebeat.com': 1, 'www.mediapost.com': 1, 'puck.news': 1, 'theweek.com': 1, 'www.detroitnews.com': 1, 'www.thetimes.com': 1, 'www.thestreet.com': 1, 'www.standard.co.uk': 1, 'mg.co.za': 1, 'hst.mit.edu': 1, 'www.thehindubusinessline.com': 1, 'www.theguardian.com': 1, 'community.nasscom.in': 1, 'boardofdirectorssalary.com': 1, 'pitchbook.com': 1, 'emailtheboss.org': 1, 'bollysuperstar.com': 1, 'www.wuft.org': 1, 'www.wboi.org': 1, 'www.investopedia.com': 1, 'www.thurrott.com': 1, 'www.xatakandroid.com': 1, 'www.aol.com': 1, 'tech.slashdot.org': 1, 'analyticsindiamag.com': 1, 'www.medianama.com': 1, 'www.goodreturns.in': 1, 'winbuzzer.com': 1, 'it.wikipedia.org': 1, 'www.punto-informatico.it': 1, 'www.bollywoodshaadis.com': 1, 'www.siliconvalley.com': 1, 'news.lankasri.com': 1, 'www.dnaindia.com': 1, 'www.indiaherald.com': 1, 'www.wwno.org': 1, 'www.northcountrypublicradio.org': 1, 'www.thehansindia.com': 1, 'www.wvxu.org': 1, 'www.nhpr.org': 1, 'www.channelnewsasia.com': 1, 'www.businessinsider.in': 1, 'bgr.com': 1}) /content/drive/MyDrive/OSINT_All/main.py:139: **UserWarning: set_ticklabels() should only be used with a fixed number of ticks, i.e. after set_ticks() or using a FixedLocator.** ax.set_xticklabels(categories, rotation=45, ha='right') All Metadata Rows (first 5): [{'Result ID': 'Hasil #1', 'Emails': 'Tidak ada', 'Phone Numbers': 'Tidak ada', 'Social Media': 'Tidak ada', 'PERSON': 'Sundar', 'NORP': 'Tidak ada', 'FAC': 'Tidak ada', 'ORG': 'Google', 'GPE': 'Tidak ada', 'LOC': 'Tidak ada', 'PRODUCT': 'Tidak ada', 'EVENT': 'Tidak ada', 'WORK_OF_ART': 'Tidak ada', 'LAW': 'Tidak ada', 'LANGUAGE': 'Tidak ada', 'DATE': 'Jan 20, 2023 ...', 'TIME': 'earlier today', 'PERCENT': 'Tidak ada', 'MONEY': 'Tidak ada', 'QUANTITY': 'Tidak ada', 'ORDINAL': 'Tidak ada', 'CARDINAL': 'Tidak ada'}, {'Result ID': 'Hasil #2', 'Emails': 'Tidak ada', 'Phone Numbers': 'Tidak ada',

'Social Media': 'Tidak ada', 'PERSON': "Sundar Pichai's", 'NORP': 'Tidak ada', 'FAC': 'Tidak ada', 'ORG': 'LinkedIn', 'GPE': 'Tidak ada', 'LOC': 'Tidak ada', 'PRODUCT': 'Tidak ada', 'EVENT': 'Tidak ada', 'WORK_OF_ART': 'Tidak ada', 'LAW': 'Tidak ada', 'LANGUAGE': 'Tidak ada', 'DATE': 'Tidak ada', 'TIME': 'Tidak ada', 'PERCENT': 'Tidak ada', 'MONEY': 'Tidak ada', 'QUANTITY': 'Tidak ada', 'ORDINAL': 'Tidak ada', 'CARDINAL': '1 billion'}, {'Result ID': 'Hasil #3', 'Emails': 'Tidak ada', 'Phone Numbers': 'Tidak ada', 'Social Media': 'Tidak ada', 'PERSON': 'Sundar Pichai', 'NORP': 'Tidak ada', 'FAC': 'Tidak ada', 'ORG': 'Tidak ada', 'GPE': 'Tidak ada', 'LOC': 'Tidak ada', 'PRODUCT': 'Tidak ada', 'EVENT': 'Tidak ada', 'WORK_OF_ART': 'Tidak ada', 'LAW': 'Tidak ada', 'LANGUAGE': 'Tidak ada', 'DATE': 'Mar 13, 2023 ...', 'TIME': 'Tidak ada', 'PERCENT': 'Tidak ada', 'MONEY': 'Tidak ada', 'QUANTITY': 'Tidak ada', 'ORDINAL': 'Tidak ada', 'CARDINAL': 'Tidak ada'}, {'Result ID': 'Hasil #4', 'Emails': 'Tidak ada', 'Phone Numbers': 'Tidak ada', 'Social Media': 'Tidak ada', 'PERSON': 'Tidak ada', 'NORP': 'Tidak ada', 'FAC': 'Tidak ada', 'ORG': 'Google', 'GPE': 'Tidak ada', 'LOC': 'Tidak ada', 'PRODUCT': 'Tidak ada', 'EVENT': 'Tidak ada', 'WORK_OF_ART': 'Tidak ada', 'LAW': 'Tidak ada', 'LANGUAGE': 'Tidak ada', 'DATE': 'Jan 20, 2023 ..., 6 months, 2 weeks, every year, 6 months', 'TIME': 'Tidak ada', 'PERCENT': 'Tidak ada', 'MONEY': 'Tidak ada', 'QUANTITY': 'Tidak ada', 'ORDINAL': 'Tidak ada', 'CARDINAL': 'Tidak ada'}, {'Result ID': 'Hasil #5', 'Emails': 'Tidak ada', 'Phone Numbers': 'Tidak ada', 'Social Media': 'Tidak ada', 'PERSON': 'Sundar Pichai', 'NORP': 'Tidak ada', 'FAC': 'Tidak ada', 'ORG': 'Tidak ada', 'GPE': 'Israel', 'LOC': 'Tidak ada', 'PRODUCT': 'Tidak ada', 'EVENT': 'Tidak ada', 'WORK_OF_ART': 'Tidak ada', 'LAW': 'Tidak ada', 'LANGUAGE': 'Tidak ada', 'DATE': 'Jun 9, 2024', 'TIME': 'Tidak ada', 'PERCENT': 'Tidak ada', 'MONEY': 'Tidak ada', 'QUANTITY': 'Tidak ada', 'ORDINAL': 'Tidak ada', 'CARDINAL': 'Tidak ada'}]

Tabel 3. SEQ Tabel * ARABIC 1 Ringkasan metadata hasil pencarian terhadap entitas "Sundar Pichai", mencakup jenis entitas, jumlah kemunculan, serta domain sumber informasi dari berbagai jenis website

Setelah proses ekstraksi metadata selesai, hasil pencarian tersebut diekspor ke dalam file berformat HTML dengan nama Results_Sundar_Pichai_AND_Email_2025-05-03_05-04-38.html bisa di lihat pada Gambar 3.3, yang tersimpan di direktori ./report/.

File ini berfungsi sebagai laporan lengkap yang mendokumentasikan semua entitas yang berhasil diidentifikasi beserta frekuensi kemunculannya, serta sumber domain tempat informasi tersebut ditemukan. Dengan laporan ini, pengguna dapat dengan mudah meninjau kembali hasil investigasi, menelusuri detail entitas, serta melakukan analisis lebih lanjut terhadap pola keterkaitan antar data.

Laporan ini juga berguna untuk kebutuhan dokumentasi, pelaporan forensik digital, atau sebagai bukti pendukung dalam kegiatan investigatif yang memerlukan akurasi dan pelacakan sumber informasi yang jelas.

Gambar 3. SEQ Gambar * ARABIC 3 Tampilan output dalam bentuk HTML

4. Tampilan Grafik Metadata
Gambar 3.4 menunjukkan visualisasi mengenai jumlah kemunculan entitas dalam metadata berdasarkan hasil ekstraksi menggunakan pendekatan Named Entity Recognition (NER). Grafik ini memberikan gambaran menyeluruh tentang kategori entitas apa saja yang paling sering muncul dalam kumpulan metadata yang dianalisis.

Secara keseluruhan, entitas dengan jumlah kemunculan terbanyak adalah "ORG" atau organisasi, dengan total 463 kemunculan. Hal ini menunjukkan bahwa metadata sangat banyak menyebutkan nama-nama lembaga, perusahaan, atau institusi, menandakan bahwa aspek institusional menjadi fokus utama dalam konten metadata tersebut. Disusul oleh entitas "PERSON" sebanyak 381 kemunculan, yang mengindikasikan banyaknya penyebutan nama individu dalam metadata. Sementara itu, entitas "DATE" muncul sebanyak 355 kali, memperlihatkan bahwa informasi terkait waktu, seperti tanggal kejadian atau penerbitan, juga sangat dominan.

Entitas lain yang juga banyak ditemukan antara lain "PRODUCT" sebanyak 75 kali dan "GPE" atau Geopolitical Entity sebanyak 57 kali, yang mencerminkan banyaknya penyebutan nama produk maupun wilayah geografis seperti negara atau kota. Sementara itu, "CARDINAL" dengan 52 kemunculan menunjukkan banyaknya informasi berbentuk angka kuantitatif yang bukan merupakan nilai mata uang, tanggal, maupun urutan.

Kategori "MONEY", "PERCENT", dan "ORDINAL" masing-masing muncul sebanyak 25, 13, dan 13 kali. Ini mengindikasikan adanya keberadaan informasi terkait nilai uang, persentase, dan urutan dalam metadata. Sementara itu, kategori lain seperti "FAC", "NORP", "LAW", "LANGUAGE", dan "WORK_OF_ART" hanya muncul satu hingga dua kali, bahkan beberapa kategori seperti "LOC" dan "EVENT" tidak muncul sama sekali. Hal ini menunjukkan bahwa informasi terkait lokasi umum, peristiwa, karya seni, hukum, dan bahasa tidak banyak terwakili dalam metadata yang dianalisis.

Selain entitas yang terdeteksi melalui model NER, grafik ini juga memuat informasi dari hasil ekstraksi elemen-elemen eksplisit seperti alamat email (14 kali), nomor telepon (4 kali), dan akun media sosial (1 kali). Meskipun jumlahnya relatif kecil, keberadaan elemen-elemen ini tetap menunjukkan bahwa metadata mengandung informasi identitas digital yang berpotensi penting dalam konteks analisis keamanan data maupun privasi.

Visualisasi ini juga dilengkapi dengan dua tombol interaktif, yaitu “Lihat Detail Metadata” dan “Lihat Semua Metadata”. Tombol Lihat Detail Metadata memungkinkan pengguna untuk melihat informasi metadata secara lebih rinci berdasarkan entitas tertentu, sedangkan tombol Lihat Semua Metadata menampilkan seluruh hasil ekstraksi yang telah dilakukan secara menyeluruh. Keberadaan fitur interaktif ini meningkatkan keterbacaan dan fungsionalitas laporan, terutama dalam eksplorasi data yang kompleks dan berskala besar.

Gambar 3. SEQ Gambar * ARABIC 4 Visualisasi jumlah kemunculan entitas dalam metadata.
5. Tampilan Grafik Jumlah Domain Dalam Pencarian
Memperlihatkan visualisasi jumlah kemunculan domain dalam hasil pencarian metadata menggunakan tool OSINT. Grafik ini berfungsi untuk menunjukkan seberapa sering setiap domain muncul sebagai sumber informasi selama proses penelusuran dilakukan. Dari grafik terlihat bahwa sebagian besar domain memiliki frekuensi kemunculan yang rendah (1-4 kali), namun ada beberapa domain yang mendominasi. Salah satu domain dengan jumlah tertinggi adalah www.w3.org yang muncul sebanyak 44 kali, menunjukkan tingginya referensi ke situs resmi World Wide Web Consortium (W3C) yang banyak digunakan dalam struktur metadata web.

Selanjutnya, domain schema.org muncul sebanyak 20 kali, yang menandakan bahwa banyak metadata merujuk pada skema standar untuk struktur data terformat yang digunakan oleh mesin pencari. Domain www.loc.gov, yang merupakan situs resmi Library of Congress, juga muncul sebanyak 10 kali, memperlihatkan kontribusi lembaga resmi dalam struktur metadata yang dianalisis.

Beberapa domain lainnya seperti prismstandard.org, dublincore.org, dan urlebird.com juga muncul dengan frekuensi yang cukup mencolok, menandakan adanya variasi dalam sumber metadata, baik dari standar internasional maupun dari media sosial atau sumber publik daring lainnya.

Dengan demikian, grafik ini memberikan gambaran bahwa hasil pencarian metadata bersumber dari berbagai domain, namun tetap terpusat pada beberapa domain utama yang menjadi standar rujukan dalam penyusunan metadata digital. Analisis ini penting dalam memahami sebaran sumber informasi serta otoritas domain yang paling sering dijadikan acuan dalam metadata hasil pencarian.

Gambar 3. SEQ Gambar * ARABIC 5 Jumlah Kemunculan Setiap Domain dalam Metadata Hasil Pencarian.

6. Tampilan Grafik Kata-kata Umum
Grafik pada gambar 3.6 yang ditampilkan memperlihatkan sepuluh kata yang paling sering muncul dalam deskripsi hasil pencarian, dengan sumbu horizontal menunjukkan kata-kata tersebut dan sumbu vertikal menunjukkan frekuensi kemunculannya. Terlihat bahwa kata “the” menempati posisi teratas dengan frekuensi tertinggi, diikuti secara berurutan oleh “pichai”, “google”, dan “sundar”. Setelah itu, frekuensi kata menurun cukup tajam pada kata “to”, lalu semakin menurun pada kata-kata berikutnya seperti “in”, “ceo”, “email”, “of”, dan “a”.

Garis merah yang menghubungkan titik-titik pada grafik membantu memperjelas pola penurunan frekuensi dari kata yang paling sering hingga yang paling jarang di antara sepuluh besar. Pola ini menggambarkan bahwa kata-kata seperti “the”, “to”, “in”, “of”, dan “a” merupakan kata umum dalam bahasa Inggris yang memang sering muncul dalam berbagai teks, sedangkan kata-kata seperti “pichai”, “google”, “sundar”, “ceo”, dan “email” adalah kata kunci yang relevan dengan topik pencarian, kemungkinan besar terkait dengan figur Sundar Pichai dan Google. Grafik ini secara visual memudahkan kita dalam memahami distribusi dan dominasi kata-kata tertentu dalam kumpulan data deskripsi hasil pencarian yang dianalisis

Gambar 3. SEQ Gambar * ARABIC 6 Grafik Sepuluh Kata yang Paling Sering Muncul dalam Deskripsi Hasil Pencarian.
7. Tampilan Grafik Jumlah Karakter dan Digit Metadata
Pada gambar 3.7, yang ditampilkan adalah sebuah diagram lingkaran (pie chart) yang menggambarkan proporsi antara jumlah karakter dan digit yang terdapat dalam metadata. Diagram ini secara visual membagi metadata menjadi dua bagian utama, yaitu karakter (ditampilkan dengan warna biru) dan digit (ditampilkan dengan warna hijau).

Dari diagram tersebut, terlihat bahwa karakter mendominasi isi metadata dengan persentase sebesar 93%. Artinya, sebagian besar data yang terdapat dalam metadata berupa huruf atau simbol non-numerik. Sementara itu, digit hanya mengambil porsi sebesar 7% dari keseluruhan metadata, yang berarti bagian data yang berupa angka atau simbol numerik sangat sedikit jika dibandingkan dengan karakter.

Visualisasi ini memberikan gambaran yang sangat jelas mengenai komposisi metadata yang dianalisis. Dengan dominasi karakter yang sangat tinggi, dapat disimpulkan bahwa data metadata yang digunakan lebih banyak berisi informasi berbentuk teks daripada angka. Informasi ini penting, terutama jika ingin memahami struktur dan jenis data yang terkandung dalam metadata tersebut, misalnya untuk keperluan analisis lebih lanjut atau pengolahan data.

Gambar 3. SEQ Gambar * ARABIC 7 Menampilkan Diagram Lingkaran yang Menunjukkan Proporsi Antara Karakter dan Digit Dalam Metadata.
8. Tampilan Grafik Jumlah Email dan Nomor Telpon
Visualisasi yang ditampilkan pada gambar 3.8 memperlihatkan perbandingan jumlah email dan nomor telepon yang ditemukan dalam data yang dianalisis. Pada sumbu horizontal (kategori), terdapat dua kategori utama, yaitu “Emails” dan “Phone Numbers”. Sementara itu, sumbu vertikal menunjukkan jumlah temuan pada masing-masing kategori.

Dari grafik tersebut, terlihat bahwa jumlah email yang ditemukan jauh lebih banyak dibandingkan dengan nomor telepon. Garis berwarna ungu yang merepresentasikan email menunjukkan angka yang cukup tinggi, yaitu sebanyak 14 email, sedangkan garis berwarna hijau yang merepresentasikan nomor telepon hanya menunjukkan angka 4. Pola garis yang saling berpotongan ini memperjelas perbedaan signifikan antara kedua kategori tersebut.

Secara visual, grafik ini memberikan gambaran yang jelas bahwa data yang dianalisis lebih banyak mengandung informasi berupa alamat email daripada nomor telepon. Hal ini bisa menjadi indikasi bahwa email lebih sering dicantumkan atau lebih mudah terdeteksi dalam sumber data yang digunakan, dibandingkan dengan nomor telepon. Grafik ini efektif dalam menunjukkan distribusi dan kecenderungan jenis data kontak yang ditemukan selama proses analisis.

Gambar 3. SEQ Gambar * ARABIC 8 Grafik visualisasi Jumlah Email dan Nomor Telpon.
9. Tampilan Grafik Network Graph Visualization
Tampilan visualisasi Network Graph Visualization pada gambar 3.9 merupakan visualisasi network graph yang secara khusus dirancang untuk menggambarkan pola hubungan antara hasil pencarian dengan berbagai entitas yang ditemukan dalam data. Setiap titik atau node pada grafik ini merepresentasikan sebuah entitas, seperti nama, institusi, alamat email, atau informasi relevan lainnya yang berhasil diekstraksi dari hasil pencarian. Sementara itu, garis-garis yang menghubungkan antar node, atau disebut juga edge, menunjukkan adanya keterkaitan atau relasi antara dua entitas yang saling berhubungan dalam konteks data yang dianalisis.

Secara visual, dapat diamati bahwa terdapat dua node utama di bagian tengah grafik yang berukuran lebih besar dan diberi warna berbeda dari node-node lainnya. Kedua node ini berfungsi sebagai pusat jaringan (central nodes), yang berarti mereka memiliki jumlah koneksi (degree) paling banyak dibandingkan dengan node-node lainnya. Hal ini menandakan bahwa kedua entitas tersebut memiliki peran yang sangat penting dan menjadi penghubung utama dalam jaringan data yang divisualisasikan. Node-node pusat ini biasanya merepresentasikan entitas yang paling sering muncul atau

paling banyak terhubung dengan entitas lain, misalnya nama tokoh utama, institusi besar, atau istilah kunci yang sering disebutkan dalam hasil pencarian.

Node-node lainnya tersebar mengelilingi pusat jaringan dengan ukuran dan warna yang bervariasi. Variasi warna dan ukuran ini biasanya merepresentasikan kategori, jenis entitas, atau tingkat keterhubungan yang berbeda. Semakin besar ukuran sebuah node, semakin tinggi pula tingkat keterhubungannya dengan node lain, sedangkan warna dapat digunakan untuk membedakan jenis entitas atau kelompok tertentu dalam data.

Garis-garis yang menghubungkan antar node memperlihatkan pola interaksi atau relasi yang kompleks di antara entitas-entitas tersebut. Semakin banyak garis yang terhubung ke sebuah node, semakin sentral peran node tersebut dalam jaringan. Pola hubungan yang terbentuk dapat membantu kita memahami struktur data secara keseluruhan, mengidentifikasi kelompok-kelompok entitas yang saling berkaitan (clustering), serta menemukan entitas kunci yang berfungsi sebagai penghubung utama (hub) dalam jaringan.

Secara keseluruhan, network graph visualization ini sangat efektif untuk memberikan gambaran menyeluruh mengenai bagaimana data hasil pencarian saling terhubung satu sama lain. Visualisasi seperti ini sangat bermanfaat dalam analisis data relasional, pemetaan jejaring sosial, serta identifikasi pusat-pusat informasi yang memiliki pengaruh besar dalam suatu sistem. Dengan memahami pola hubungan yang divisualisasikan dalam grafik ini, analis data dapat mengambil keputusan yang lebih tepat dalam proses pengolahan, interpretasi, maupun pengembangan strategi berdasarkan data yang tersedia.

Gambar 3. SEQ Gambar * ARABIC 9 Visualisasi Network Graph yang memperlihatkan pola hubungan antara hasil pencarian dan berbagai Entitas Terkait, di mana dua Node Utama di Tengah menjadi pusat Keterhubungan Dalam Jaringan.

10. Langkah Identifikasi dan Mitigasi
Langkah identifikasi pada tahap ini bertujuan untuk menentukan dan menyaring data email yang diperoleh dari hasil ekstraksi menggunakan tool OSINT, yang disimpan dalam file Results_Sundar_Pichai_AND_Email_2025-05-03_05-04-38.html. Data email tersebut diambil melalui antarmuka interaktif, khususnya dari tombol “Lihat Detail Entitas”, yang ketika diklik akan menampilkan pop-up berisi hasil ringkasan entitas dari proses Named Entity Recognition (NER). Ringkasan ini mencakup berbagai informasi penting seperti alamat email, nomor telepon, dan akun media sosial.

Selain tombol tersebut, terdapat juga fitur “Lihat Semua Metadata” yang menyajikan seluruh hasil metadata secara lengkap. Namun karena data yang ditampilkan melalui fitur ini bersifat sangat luas dan tidak tersaring, maka untuk tujuan identifikasi alamat email secara efisien dan terfokus, data diambil hanya dari tampilan ringkas tombol “Lihat Detail Entitas”. Pemilihan ini juga memudahkan proses lanjutan, karena informasi yang ditampilkan sudah tersusun berdasarkan kategori entitas yang relevan.

Setelah email berhasil diidentifikasi, langkah berikutnya adalah melakukan pemeriksaan keamanan terhadap email-email tersebut. Pemeriksaan ini dilakukan menggunakan layanan Have I Been Pwned (HIBP), sebagaimana ditampilkan pada Gambar 3.10. Proses pengecekan ini bertujuan untuk mengetahui apakah alamat email tersebut pernah terlibat dalam insiden kebocoran data, baik akibat peretasan, pelanggaran keamanan sistem, maupun praktik pengumpulan data ilegal.

Informasi yang diperoleh dari hasil pengecekan ini menjadi dasar awal untuk menilai tingkat risiko terhadap penyalahgunaan identitas digital. Dengan mengetahui status keterlibatan email dalam insiden kebocoran, pengguna atau peneliti dapat melakukan analisis lebih lanjut terkait perlindungan data pribadi dan potensi ancaman siber.

Gambar 3. SEQ Gambar * ARABIC 10 Pemeriksaan keamanan alamat email menggunakan layanan Have I Been Pwned untuk mendeteksi apakah email terdaftar pernah terlibat dalam insiden kebocoran data

Tahap selanjutnya setelah proses identifikasi dan pengecekan adalah tahap mitigasi, yakni serangkaian tindakan yang dirancang untuk mengurangi potensi risiko keamanan yang mungkin timbul dari data yang telah teridentifikasi. Langkah-langkah mitigasi ini disusun secara selektif berdasarkan status hasil pemeriksaan terhadap alamat email apakah terindikasi mengalami kebocoran data atau tidak.

Mitigasi dilakukan dengan mempertimbangkan sejauh mana email tersebut terekspos dalam insiden kebocoran, serta bagaimana email itu digunakan dalam konteks digital. Jika email diketahui pernah terlibat dalam pelanggaran keamanan, maka tindakan lanjut diperlukan untuk meminimalisir dampak, melindungi privasi pemiliknya, dan mencegah penyalahgunaan di masa mendatang. Sebaliknya, jika tidak ditemukan indikasi kebocoran, maka langkah preventif tetap dilakukan dalam bentuk penyimpanan data secara terenkripsi dan pemantauan berkala.

Langkah-langkah mitigasi berdasarkan hasil pengecekan email dirangkum dalam Tabel 3.1 berikut:

Status Hasil Pengecekan Langkah Mitigasi yang Dilakukan
Email Terindikasi Bocor - Memberikan peringatan kepada pemilik email (jika dapat dihubungi). - Menyarankan untuk segera mengganti kata sandi. - Menganjurkan penggunaan autentikasi dua faktor (2FA). - Menyimpan email dalam daftar observasi untuk pemantauan lanjutan. - Mengenkripsi hasil penyimpanan data dengan menggunakan enkripsi AES 256 atau PGP untuk keamanan data secara lokal agar tidak mudah diakses oleh pihak tidak berwenang.
Email Tidak Terindikasi Bocor - Menandai email sebagai "aman sementara". - Menyimpan data secara terenkripsi sebagai langkah preventif. - Tidak dilakukan tindakan lanjut kecuali ditemukan dalam pencarian berikutnya. - Dicatat sebagai referensi perbandingan status keamanan di masa mendatang.

Tabel 3. SEQ Tabel * ARABIC 2 Langkah Mitigasi Berdasarkan Status Kebocoran Email
4.
5. VII. Simpulan

Berdasarkan hasil penelitian dan analisis yang telah dilakukan, dapat disimpulkan bahwa penggunaan teknik Open Source Intelligence (OSINT) sebagai sumber data, dikombinasikan dengan metode footprinting sebagai proses pengumpulan informasi, terbukti cukup efektif dalam pelacakan jejak digital suatu entitas. Jejak digital yang terkumpul berperan penting dalam memperkaya informasi mengenai target yang dianalisis, baik berupa identitas, aktivitas, maupun potensi risiko keamanan yang berkaitan.

Dalam implementasinya, artikel ini mengembangkan sebuah alat bantu (tools) berbasis Command Line Interface (CLI) dengan menggunakan bahasa pemrograman Python. Tool ini memanfaatkan teknik pembelajaran mesin (Machine Learning), khususnya Named Entity Recognition (NER), untuk mengekstraksi entitas dari metadata. Selain itu, digunakan juga teknik regular expression untuk mengekstraksi informasi eksplisit seperti alamat email dan nomor telepon. Output yang dihasilkan berupa file HTML yang tidak hanya memuat data mentah, tetapi juga dilengkapi dengan berbagai visualisasi grafik yang mempermudah proses analisis dan interpretasi data oleh pengguna.

Lebih lanjut, sebagai bagian dari proses identifikasi dan mitigasi risiko, email yang ditemukan dianalisis menggunakan layanan Have I Been Pwned (HIBP) untuk memverifikasi apakah email tersebut pernah terlibat dalam kebocoran data. Langkah ini memberikan nilai tambah signifikan dalam konteks keamanan siber, karena tidak hanya mendeteksi potensi ancaman, tetapi juga memfasilitasi langkah-langkah mitigasi secara tepat dan terukur.

Secara keseluruhan, integrasi antara metode OSINT, footprinting, pemrosesan metadata berbasis NER, serta validasi data menggunakan HIBP menjadikan sistem ini cukup efektif dalam mendukung proses investigasi digital. Pendekatan ini dapat menjadi acuan untuk pengembangan alat forensik digital atau keamanan siber di masa mendatang, terutama dalam konteks pengumpulan dan analisis informasi terbuka secara otomatis dan sistematis.
6.
7. Referensi

[1] “2025 Data Breach Investigations Report,” Verizon Business. Accessed: Apr. 28, 2025. [Online]. Available: https://www.verizon.com/business/resources/reports/dbir/

[2] “Open Source Intelligence Techniques,” Guide books. Accessed: Apr. 28, 2025. [Online]. Available: https://dl.acm.org/doi/abs/10.5555/3033260

[3] M. F. Alby, I. F. Ruslan, and M. L. Muharman, “Information Security Test on Websites and Social Media Using Footprinting Method,” in Proceedings of the 8th International Conference on Industrial and Business Engineering, in ICIBE ’ **22. New York, NY, USA: Association for Computing Machinery,** Jan. 2023, pp. 521-525. doi: 10.1145/3568834.3568868.

[4] X. Schmitt, S. **Kubler**, J. **Robert**, M. **Papadakis**, and Y. **LeTraon**, “ A **Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate,” in** 2019 Sixth **International Conference on Social Networks Analysis, Management and Security** (SNAMS), Oct. **2019, pp.** 338-343. doi: 10.1109/ SNAMS. 2019.8931850.

[5] “OSINT in the Context of Cyber-Security | SpringerLink.” Accessed: Apr. 28, 2025. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-47671-1_14

[6] “CSS, Bootstrap, & Responsive Design | SpringerLink.” Accessed: Apr. 28, 2025. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4842-2044-3_6

[7] “Discovering Personal Data Security Issues: Insights from ‘Have I Been Pwned’ | SpringerLink.” Accessed: Apr. 28, 2025. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-70906-7_22

[8] “ **TATA KELOLA PERLINDUNGAN DATA PRIBADI DI ERA METAVERSE (TELAAH YURIDIS UNDANG-UNDANG PERLINDUNGAN DATA** PRIBADI) | Sulistianingsih | Masalah-Masalah **Hukum.”** Accessed: Apr. 28, 2025. [Online]. Available: https://ejournal.undip.ac.id/index.php/mmh/article/view/51319

[9] K. Shye Lianq and V. Selvarajah, “Footprinting and Reconnaissance: Impact and Risks,” **in 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE),** Apr. **2022, pp. 1-5. doi: 10.1109/** ICDCECE53908.2022.9793237.

[10] S. Harris, “All-in-one CISSP exam guide,” (No Title), Accessed: Apr. 28, 2025. [Online]. Available: https://cir.nii.ac.jp/crid/1130000798088792832

[11] M. Y. Samad, B. K. Ningtiyas, Fiqih, F. Rosny, and D. A. Permatasari, “Anticipating Cyber Espionage: Open Source Intelligence (OSINT) Investigation and Cyber Counterintelligence,” JSRCS, vol. 5, no. 2, pp. 167-184, Nov. 2024, doi: 10.31599/288ab341.

[12] **I. Harouni, “The Modern Methods of Data Analysis in Social Research,”** sej, vol. 6, no. **1, pp.** 56-70, Mar. 2024, doi: **10.34118/sej.v6i1.3806.**

[13] “Proposing a New Combined Indicator for Measuring Search Engine Performance and Evaluating Google, Yahoo, DuckDuckGo, and Bing Search Engines based on Combined Indicator - Azadeh Hajian Hoseinabadi, Mehrdad CheshmehSohrabi, 2024.” Accessed: Apr. 29, 2025. [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/09610006221138579

[14] “**Application of Named Entity Recognition via Twitter on SpaCy in Indonesian (Case Study): Power Failure in the Special Region of Yogyakarta)** | **Indonesian Journal of Information Systems.”** Accessed: Apr. 29, 2025. [Online]. Available: https://ojs.uajy.ac.id/index.php/IJIS/article/view/4677

[15] “Introduction to Bootstrap | SpringerLink.” Accessed: Apr. 29, 2025. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4842-6203-0_1

[16] “ **Have I Been Pwned: Check if your email has been compromised in a data breach.”** Accessed: Apr. 29, 2025. [Online]. Available: https://haveibeenpwned.com/