 doi.org/10.56706/ik.v19i3.127

Wacana Media tentang Pengungkapan Data Tanpa Izin di Indonesia: Analisis Frekuensi Kata dan Cloud

Fathan Aldi Rivai[1)]

1) *Regional Planning Science Study Program, specializing in Big Data and Spatial Modelling, Department of Soil Science and Land Resources, Faculty of Agriculture, IPB University, r.fathan.aldi@gmail.com*

Riwayat Artikel

Dikirim: 2 Jul 2025
Diterima: 12 Nov 2025
Diterbitkan: 9 Des 2025

Kata kunci:

Isu data
Kebocoran data
Korpus
Penambangan teks
Wacana media

Keywors:

Data issues
Data leaks
Corpus
Text mining
Media discourse

Abstrak (10 PT)

Pengungkapan data secara tidak sah semakin sering terjadi di Indonesia, memunculkan kekhawatiran publik terkait keamanan data pribadi. Fenomena ini menjadi perhatian media untuk membentuk narasi dan menyampaikan perspektif. Penelitian ini bertujuan untuk mengungkap bagaimana media mengonstruksi isu data melalui analisis frekuensi kata dan word cloud. Data dikumpulkan dari 212 artikel berita sepanjang tahun 2024 yang berasal dari CNN Indonesia, Detik, Kompas, dan Liputan6. Penambangan kata artikel dilakukan dan dianalisis menggunakan R untuk memperoleh kata-kata dominan yang membentuk narasi media. Hasil menunjukkan bahwa kata "bocor", "akses", "jual", dan "curi" mendominasi pemberitaan, dengan penekanan besar pada tindakan, bukan hanya aktor atau lembaga. Kata kerja muncul lebih sering dibandingkan nama tokoh maupun institusi, menunjukkan bahwa media lebih fokus pada aksi dalam pemberitaan kebocoran data. Visualisasi *word cloud* memperkuat temuan ini, dengan ukuran kata kerja yang lebih mencolok. Kontribusi penelitian ini memberikan pemahaman terhadap wacana media dalam isu kebocoran data pribadi di Indonesia, dan mengungkap kecenderungan jenis kata yang dominan dimunculkan.

***Abstract* (10 PT)**

Unauthorized data disclosures are increasingly frequent in Indonesia, raising public concerns over the security of personal information. This phenomenon has drawn media attention, shaping narratives and conveying specific perspectives. This study aims to examine how the media constructs the issue of data breaches through word frequency and word cloud analysis. A total of 212 news articles published throughout 2024 were collected from CNN Indonesia, Detik, Kompas, and Liputan6. Text mining was conducted using R to extract and analyze dominant words that form media narratives. The results reveal that words such as "leak," "access," "sell," and "steal" dominate the coverage, highlighting a strong emphasis on actions rather than actors or institutions. Verbs appear more frequently than the names of individuals or organizations, indicating that the media tends to focus on actions when reporting on data breaches. The word cloud visualization reinforces this finding, with verbs appearing in significantly larger sizes. This study contributes to a deeper understanding of media discourse on personal data breaches in Indonesia and reveals the tendency for certain types of words to dominate the narrative.

1. PENDAHULUAN

Dengan kemajuan teknologi yang pesat, pengumpulan data dalam format digital telah menjadi tren global di berbagai sektor [1]. Transformasi digital ini tidak hanya menyederhanakan proses dokumentasi dan analisis tetapi juga meningkatkan efisiensi pengambilan keputusan [2]. Di sektor pemerintahan, pendidikan, dan bisnis, digitalisasi data memungkinkan integrasi sistem yang lebih efektif dan mempercepat akses ke informasi yang dibutuhkan [3]. Di Indonesia, pengumpulan data warga negara dalam bentuk digital, seperti Kartu Tanda Penduduk (KTP), Surat Izin Mengemudi (SIM), Nomor Pokok Wajib Pajak (NPWP), dan informasi pribadi lainnya telah semakin meluas seiring dengan upaya digitalisasi layanan publik. Pemerintah Indonesia telah menerapkan berbagai sistem elektronik untuk mengefisiensikan proses administrasi dan meningkatkan efisiensi pemberian layanan publik. Salah satu inisiatif tersebut adalah adopsi Kartu Tanda Penduduk Elektronik (e-KTP), yang bertujuan untuk memperbarui data identitas kependudukan dan memodernisasi sistem administrasi sipil.

Data pribadi seperti dokumen identitas seringkali diperlukan untuk pendaftaran administratif di seluruh platform digital, termasuk aplikasi perbankan, *e-commerce*, dan layanan sosial. Meskipun digitalisasi menawarkan kemudahan, meluasnya penggunaan data pribadi menimbulkan kekhawatiran akan potensi penyalahgunaan, pelanggaran data, dan akses tanpa izin ke informasi sensitif [4]. Perkembangan ini menghadirkan tantangan baru, terutama dalam hal keamanan, privasi, dan etika dalam pengelolaan data digital [5]. Tanpa langkah-langkah keamanan siber yang memadai, risiko pelanggaran dan pengungkapan data tanpa izin akan terus meningkat. Pengungkapan tanpa izin mengacu pada tindakan mengungkap data yang seharusnya dilindungi kepada pihak yang tidak berwenang, sehingga mengancam kerahasiaan informasi [6]. Praktik ini mencakup pelanggaran data, penyalahgunaan informasi, akses dan distribusi tanpa izin, serta pengungkapan yang melanggar peraturan perlindungan data. Insiden semacam itu semakin umum terjadi di Indonesia, menempatkan negara ini di peringkat ke-14 dunia dalam peringkat pelanggaran data pada tahun 2025 [7]. Frekuensi kejadian ini telah menyebabkan liputan media rutin tentang fenomena ini di Indonesia.

Sebagai pengawas publik, media Indonesia menjadi lebih aktif dalam melaporkan pengungkapan data tanpa izin. Setiap kali terjadi pelanggaran data, media tidak hanya menyajikan fakta tetapi juga menyoroti potensi dampaknya, baik bagi individu yang terdampak maupun bagi kepercayaan publik terhadap sistem manajemen data. Laporan media seringkali menekankan perspektif organisasi, yang cenderung mengaitkan pelanggaran dengan faktor eksternal sambil meremehkan respons otoritas dan lembaga terkait [8]. Liputan juga mencakup pendapat dan komentar ahli yang menawarkan rekomendasi tentang langkah-langkah pencegahan untuk mengurangi kemungkinan pelanggaran di masa mendatang [9]. Dengan demikian, media tidak hanya menginformasikan tetapi juga secara aktif membentuk kesadaran publik tentang pentingnya perlindungan data pribadi.

Studi tentang pengaplikasian analisis *word frequency and cloud* terdahulu telah banyak diaplikasikan untuk menyelediki postingan yang berkembang di sosial media [10], [11], [12], review suatu produk [13], [14], [15] dan newspaper [16], [17]. Dengan meningkatnya jumlah pelanggaran data dan pengungkapan yang tidak sah, media memainkan peran penting dalam membentuk kesadaran publik dan memengaruhi kebijakan perlindungan data [18]. Oleh karena itu, menyelidiki wacana media tentang pengungkapan data yang tidak sah di Indonesia sangat relevan, terutama pada tahun 2024, ketika isu ini telah mendapatkan perhatian yang lebih tinggi di sektor publik dan swasta. Dengan menggunakan analisis frekuensi kata dan awan kata, studi ini bertujuan untuk memberikan representasi yang jelas tentang bagaimana media mengkonstruksi isu pengungkapan data yang tidak sah. Penelitian ini memberikan kontribusi dalam memperluas pemahaman mengenai konstruksi wacana media terkait kebocoran data pribadi di Indonesia, khususnya melalui identifikasi pola dominasi jenis kata tertentu dalam pemberitaan.

2. LANDASAN TEORI

2.1. Wacana Media

Media tidak hanya menyampaikan informasi, tetapi juga merepresentasikan realitas sosial melalui narasi, simbol, dan struktur linguistik yang dibangun secara sistematis [19]. Operating within an ideological framework, media discourse reflects power relations, politico Beroperasi dalam kerangka ideologis, wacana media mencerminkan hubungan kekuasaan, kepentingan politik-ekonomi, dan nilai-nilai hegemonik melalui pemilihan kata, konstruksi naratif, dan pembingkaian isu [20]. Akibatnya, analisis wacana media tidak hanya melibatkan aspek linguistik tetapi juga dinamika kekuasaan dan representasi sosial, menjadikannya alat hegemoni simbolik dalam teori kritis. Wacana media secara inheren bersifat performatif dan selektif, mencerminkan proses pemilihan acara, penyampaian naratif, dan kurasi sumber yang tidak netral. Dinamika ini dibentuk oleh interaksi antara struktur media, tekanan politik, kepentingan ekonomi, dan ekspektasi audiens. Akibatnya, wacana media tetap dinamis dan responsif terhadap perubahan sosial, politik, dan teknologi, yang menggambarkan bagaimana media, sebagai agen produksi pengetahuan, membentuk persepsi publik [21].

2.2. Pengungkapan Data secara Tidak Sah

Pengungkapan data yang tidak sah mengacu pada situasi di mana informasi yang seharusnya dilindungi oleh entitas yang berwenang justru terekspos atau diakses oleh individu atau entitas tanpa izin yang semestinya. Tindakan tersebut dapat diakibatkan oleh kebocoran informasi akibat kegagalan keamanan sistem atau penyalahgunaan hak akses yang sah, yang sering dikaitkan dengan serangan siber, eksploitasi kerentanan, dan kesalahan manusia dalam kebijakan dan prosedur keamanan [22]. Teknik perlindungan data digital, seperti enkripsi, kontrol akses, autentikasi, dan pencatatan audit, sangat penting untuk memastikan kerahasiaan dan integritas data. Namun, pengungkapan yang tidak sah sering terjadi karena kegagalan implementasi, kelemahan algoritma, atau kesalahan manusia [23]. Serangan seperti *Man-in-the-Middle* (MITM), *buffer overflow*, dan injeksi SQL merupakan beberapa penyebab utama pelanggaran data dalam sistem berbasis basis data [24]. Di tingkat organisasi, pelanggaran data sering kali berasal dari penegakan kebijakan keamanan yang buruk, seperti implementasi yang tidak tepat dari prinsip hak istimewa terkecil, yang dirancang untuk memberikan pengguna atau aplikasi hanya hak akses minimum yang diperlukan [25]. Memperkuat sistem keamanan siber sangat penting tidak hanya untuk mencegah pelanggaran data, tetapi juga untuk memungkinkan deteksi dini ancaman dan respons insiden yang cepat. Perangkat seperti sistem deteksi intrusi (IDS) dan pemantauan *log* sangat penting untuk mengidentifikasi aktivitas mencurigakan, namun keduanya harus terus ditingkatkan untuk mendeteksi serangan yang semakin kompleks dan merespons ancaman secara otomatis [26].

2.3. Kerangka Konseptual *Word Frequency* dan Studi Terdahulu

Frekuensi kata adalah konsep inti dalam analisis teks yang berfokus pada tingkat kemunculan kata dalam korpus atau kumpulan data. Konsep ini berakar pada teori distribusi leksikal yang diperkenalkan oleh Zipf pada tahun 1949 [27], yang menyatakan bahwa kata-kata yang muncul lebih sering dalam teks cenderung memiliki signifikansi tematik yang lebih besar. Hukum Zipf menunjukkan bahwa hanya sejumlah kecil kata yang muncul dengan frekuensi tinggi, sementara sebagian besar jarang muncul. Secara statistik, analisis frekuensi kata digunakan dalam penambangan teks untuk mengungkap pola atau hubungan tersembunyi dalam data tekstual [28]. Dengan menganalisis frekuensi kata, peneliti dapat mengidentifikasi elemen-elemen kunci dalam teks, mengeksplorasi struktur tematik, dan menentukan konsep-konsep sentral yang mendominasi wacana. Dalam aplikasi praktis, analisis frekuensi kata memungkinkan eksplorasi hubungan istilah di berbagai konteks, seperti wacana media, analisis sentimen, atau studi opini [29]. Lebih lanjut, konsep frekuensi kata mendukung teknik-teknik analisis seperti analisis kohesi tekstual dan penciptaan awan kata, yang secara visual menggambarkan keunggulan dan hubungan antara istilah-istilah dalam korpus [30]. Dengan demikian, frekuensi kata berfungsi sebagai alat penting untuk mengeksplorasi pola linguistik dan mengungkap informasi laten dalam data tekstual.

Penelitian Heryono et al. [31] mengkaji pola frekuensi kata dalam artikel linguistik yang terindeks SINTA (*Science and Technology Index*) menggunakan pendekatan *text mining* berbasis korpus. Hasilnya menunjukkan terdapat kata-kata seperti "*slang*", "*metaphors*", "corona", dan "virus", yang mencerminkan bahwa penelitian saat itu banyak berhubungan dengan konteks pandemi. Studi Roy [32] mengkaji kata-kata yang paling sering muncul pada aturan perpustakaan. Hasilnya menemukan bahwa, 'buku', diikuti oleh 'kartu', 'pinjaman', 'keanggotaan', 'online', 'fakultas', '*access*', '*research*', 'material' dan 'cadangan' adalah istilah frekuensi tinggi. Walaupun studi frekuensi kata lebih banyak diaplikasikan pada teks berbahasa inggris, namun Başaran [33] mengkaji frekuensi kata pada *Language Teaching Textbooks* dan berhasil menunjukkan bahwa pola kemunculan kosakata dalam buku ajar dapat diprioritaskan untuk diajarkan kepada pelajar asing.

2.4. Dasar Teori *Word Cloud* dan Studi Terdahulu

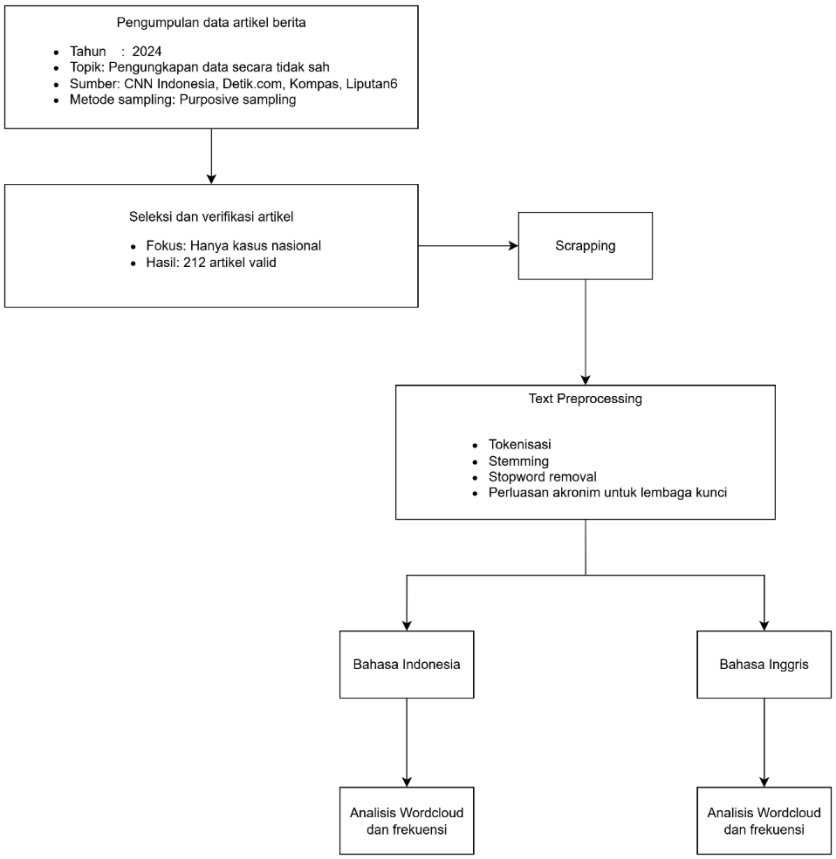
Awan kata adalah bentuk visualisasi data teks yang menyorot kata-kata yang paling sering muncul dalam sebuah teks, ditampilkan menggunakan berbagai ukuran atau warna sesuai dengan frekuensinya. Konsep ini berasal dari visualisasi informasi dan linguistik komputasional, yang berfungsi sebagai alat yang efisien untuk mengidentifikasi istilah-istilah kunci dan tema-tema menyeluruh dengan cepat [34]. Dalam kerangka semiotik, awan kata dikaitkan dengan konsep sintaksis visual, di mana susunan spasial kata-kata dirancang untuk menyampaikan makna dengan menekankan elemen-elemen yang muncul lebih sering yang mencerminkan signifikansi semantik dan saliensi kognitif [35]. Hal ini menjadikan awan kata sebagai alat yang berharga dalam analisis budaya, di mana metode komputasi digunakan untuk mengekstrak dan menafsirkan pola data, sehingga mengungkap fenomena budaya dan sosial yang mendasarinya yang tertanam dalam teks [36].

Secara kognitif, awan kata membantu mengurangi kelebihan informasi, menyederhanakan penyajian data, dan mengarahkan perhatian pada konsep-konsep inti melalui saliensi visual [37]. Konsep ini memainkan peran penting dalam memfasilitasi pemahaman pola-pola utama dalam sebuah teks. Secara linguistik, awan kata didasarkan pada analisis frekuensi kata, yang mengasumsikan bahwa frekuensi kemunculan kata dalam suatu korpus memberikan wawasan tentang struktur tematik teks dan fokus utama analisis [38]. Pendekatan ini mendukung interpretasi teks melalui distribusi leksikal, yang mana kata-kata yang sering muncul dianggap penting bagi pemahaman tematik wacana.

Studi Sasongko et al. [39] menganalisis *word cloud* yang terdapat di *science techno park* di Universitas Indonesia. Hasil analisis kata awan menunjukkan bahwa kata "teknologi", "ilmu pengetahuan," "*research*," "zona," dan "inovasi," yang merupakan elemen–elemen dari *science techno park*. Rahmadani dan Yuadi [40] melakukan analisis *word cloud* tentang Pemakzulan Yoon Seok-Yeol untuk memahami narasi politik. Hasilnya menunjukkan bahwa kata "*impeachment*," "*martial*," "*law*," "Yoon," "*party*," "*opposition*," dan "*country*" tampak lebih besar dibandingkan kata lainnya. Penelitian Sağlam [41] menunjukkan bahwa analisis *word cloud* terhadap judul artikel dapat mengungkap fokus utama dalam suatu bidang studi. Dalam studi mengenai transportasi maritim, kata-kata seperti "*port*", "*shipping*", dan "*container*" adalah yang paling dominan, Dengan demikian, pendekatan *word cloud* tidak hanya menawarkan gambaran visual yang menarik, tetapi juga mampu merepresentasikan pola diskursif yang berulang dalam teks-teks yang dianalisis.

3. METODOLOGI PENELITIAN

Studi ini terdiri dari beberapa tahap, meliputi pengumpulan data berita, penyiapan data, analisis frekuensi kata, dan visualisasi awan kata. Data bersumber dari artikel media yang diberi label "pelanggaran data", dengan kriteria artikel tersebut secara spesifik membahas masalah pelanggaran data yang terjadi di Indonesia. Artikel yang membahas pelanggaran data di negara lain tidak diikutsertakan. Sumber media yang digunakan antara lain CNN Indonesia, Detik, Kompas, dan Liputan6. Hanya artikel yang diterbitkan antara Januari dan Desember 2024 yang dipertimbangkan. Selama periode ini, total 212 artikel berita berhasil dikumpulkan. Persiapan data, pembuatan awan kata, dan analisis frekuensi kata dilakukan menggunakan R Studio [42]. Diagram alur penelitian disajikan pada Gambar 1.



Gambar 1. Research flowchart

3.1. Pengumpulan Data Berita

Langkah pertama penelitian ini melibatkan pengumpulan data berupa artikel berita yang membahas pengungkapan data pribadi tanpa izin di Indonesia sepanjang tahun 2024. Proses ini dilakukan dengan meninjau dan mengambil artikel-artikel kredibel yang memuat tag "pelanggaran data" dari media seperti CNN Indonesia, Detik, Kompas, dan Liputan6. Meskipun sumber-sumber kredibel ini juga melaporkan insiden pelanggaran data internasional, proses verifikasi dilakukan untuk menyaring dan memastikan bahwa hanya kasus-kasus domestik yang dipilih untuk dianalisis. Oleh karena itu, metode pengambilan sampel purposif digunakan, dengan fokus khusus pada konten yang ditargetkan. Artikel-artikel yang telah dikurasi kemudian dipersiapkan untuk analisis frekuensi kata dan *word cloud* selanjutnya.

3.2. Preparasi Data

Sebanyak 212 artikel berita di-*scraping* menggunakan R, memanfaatkan pustaka *rvest* [29] untuk mengekstrak elemen HTML dari halaman web. Langkah selanjutnya melibatkan tokenisasi, yang dilakukan menggunakan pustaka *tidytext* [43], yang memecah teks menjadi kata-kata individual (token). Ini diikuti oleh stemming dalam bahasa Indonesia, yang bertujuan untuk mengubah kata-kata menjadi bentuk akarnya. Kata-kata yang tidak relevan dan tidak penting, seperti konjungsi dan kata kerja yang tidak terkait langsung dengan masalah penggunaan data yang tidak sah, dihilangkan selama proses pembersihan. Selain itu, langkah penggantian dilakukan untuk mengklarifikasi istilah dengan memperluas akronim lembaga-lembaga kunci yang sering disebutkan dalam artikel. Akhirnya, teks diterjemahkan ke dalam bahasa Inggris sebagai persiapan untuk analisis lebih lanjut, termasuk frekuensi kata dan pembangkitan awan kata, yang keduanya dilakukan dalam bahasa Indonesia dan Inggris.

3.1. Analisis *Word Frequency*

Analisis frekuensi kata dilakukan untuk mengidentifikasi kata-kata yang paling sering muncul dalam kumpulan artikel berita. Proses ini dimulai dengan menghitung jumlah kemunculan setiap kata dalam dataset yang telah diproses sebelumnya menggunakan fungsi count dari pustaka *dplyr* [44]. Secara matematis, frekuensi kata w_i dalam teks dihitung menggunakan Persamaan (1).

$$f(w_i) = \sum_{j=1}^{N} 1(w_i, t_j) \quad (1)$$

Di mana $f(w_i)$ merepresentasikan frekuensi kata w_i, N menunjukkan jumlah artikel, dan $1(w_i,t_j)$ 128ndica fungsi 128ndicator yang menghasilkan 1 jika kata w_i muncul dalam artikel t_j, dan 0 jika tidak. Setelah menghitung frekuensi, kata-kata difilter menggunakan kondisi $f(w_i) > 49$, untuk mengecualikan kata-kata yang jarang muncul dan berkontribusi minimal terhadap analisis. Walaupun penelitian-penelitian terdahulu tidak menentukan ambang batas [45], [46], [47], dalam penelitian ini ditentukan untuk mengurangi *noise* dari kata-kata berfrekuensi rendah yang cenderung tidak memiliki makna analitis signifikan, serta memastikan bahwa kata-kata yang dianalisis memiliki relevansi tematik yang cukup kuat dalam konteks keseluruhan data. Ambang batas ini juga membantu meningkatkan keandalan hasil analisis dan visualisasi, seperti *word cloud*, dengan menekankan kata-kata yang benar-benar merepresentasikan pola umum dalam korpus. Probabilitas kemunculan setiap kata w_i dalam teks kemudian dihitung menggunakan Persamaan (2).

$$P(w_i) = \frac{f(w_i)}{T} \quad (2)$$

Di mana T mewakili jumlah total kata dalam teks. Analisis ini memungkinkan identifikasi kata-kata dominan dan relevan secara kontekstual dalam 128umpulan data.

3.1. Analisis *Word Cloud*

Visualisasi awan kata dilakukan menggunakan pustaka *wordcloud2* [48]. Ukuran setiap kata dalam awan kata mencerminkan frekuensinya, dengan kata-kata yang lebih sering muncul ditampilkan dalam font yang lebih besar. Metode ini dimulai dengan tabel frekuensi kata yang difilter. Setiap kata kemudian dipetakan ke ukuran dan warna tertentu berdasarkan frekuensinya, menggunakan Persamaan (3).

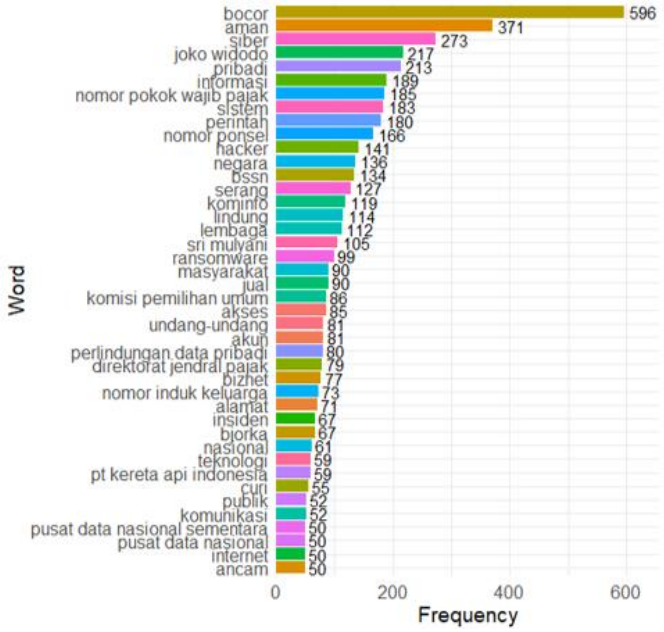
$$u(w_i) = \alpha \cdot f(w_i) + \beta \quad (3)$$

Di mana $u(w_i)$ menunjukkan ukuran visual kata w_i, $f(w_i)$ adalah frekuensi kata, a adalah faktor skala yang menyesuaikan ukuran kata, dan β adalah konstanta yang menentukan ukuran kata minimum yang terlihat. Kata-kata dengan frekuensi yang lebih tinggi tampak lebih besar, sedangkan kata-kata dengan frekuensi yang lebih rendah ditampilkan dalam ukuran yang lebih kecil. Faktor skala a mengontrol sejauh mana frekuensi memengaruhi ukuran visual, sedangkan β memastikan bahwa semua kata tetap terlihat terlepas dari frekuensinya.

4. HASIL DAN PEMBAHASAN

Isu pengungkapan data tanpa izin muncul sebagai kontroversi besar dan menjadi salah satu topik paling viral di tahun 2024. Kasus serupa terus bermunculan bahkan dua tahun setelah era "normal baru", periode yang ditandai dengan pesatnya digitalisasi pasca-pembatasan aktivitas selama pandemi. Berdasarkan hasil analisis frekuensi kata, kata yang paling sering muncul adalah "*leaking*", sebanyak 596 kali. Kata-kata dengan frekuensi terendah, yang difilter dengan ambang batas minimal 50 kemunculan, meliputi "pusat data nasional sementara" atau "Pusat Data Nasional Sementara" (PDNS), "pusat data nasional" atau "Pusat Data Nasional (PDN), "internet", dan "*threat*". Frekuensi kata dalam bahasa Indonesia dan bahasa Inggris yang diperoleh dari proses penambangan teks disajikan pada Gambar 2. Dominasi "*leaking*" mencerminkan kenyataan bahwa

sebagian besar pengungkapan data tanpa izin melibatkan kebocoran data. Kebocoran ini terjadi baik di perusahaan swasta maupun lembaga pemerintah. Di antara hasil filter dengan minimal 50 kemunculan, hanya Biznet dan PT Kereta Api Indonesia (KAI) yang muncul sebagai perusahaan, sementara sisanya mewakili sektor pemerintah.



Gambar 2. *Word frequency result*, (A) *Indonesia version*, (B) English version

Dugaan kebocoran data pribadi milik sekitar 380.000 pelanggan Biznet di web gelap telah menjadi perhatian serius, terutama terkait dengan masalah perlindungan data dan keamanan siber di Indonesia. Web Gelap adalah bagian kecil dari Web Dalam, yang merupakan segmen tersembunyi dari internet yang tidak dapat diindeks oleh mesin pencari biasa seperti Google atau Bing, tetapi dapat diakses melalui Tor (*The Onion Router*), Freenet, I2P, dan JonDonym. Sementara Web Dalam mencakup berbagai jenis konten non-publik, Web Gelap sering dikaitkan dengan aktivitas ilegal dan konten tersembunyi. Sekitar 57% penggunaannya melibatkan aktivitas kriminal dan konten ilegal, yang umumnya digunakan untuk transaksi ilegal [49]. Laporan awal menunjukkan bahwa insiden itu diduga dilakukan oleh seseorang dari dalam perusahaan yang tidak setuju dengan Kebijakan Penggunaan Wajar (*Fair Usage Policy*/FUP), yang merupakan kebijakan Biznet untuk membatasi penggunaan internet. FPU ini membatasi akses internet bagi pelanggan, agar tidak menggunakan kuota internet 1TB per bulan. Data yang bocor mencakup informasi sensitif seperti nomor ID nasional, nomor pajak, nomor telepon, dan alamat pelanggan. Situasi ini telah menimbulkan kekhawatiran publik tentang seberapa aman data pribadi sebenarnya. Menanggapi hal ini, Biznet kemudian melakukan proses investigasi untuk membuktikan kebenaran masalah ini dengan melibatkan pihak berwenang.

Kasus dugaan kebocoran data di PT Kereta Api Indonesia, yang diklaim oleh kelompok ransomware Stormous, masih dalam penyelidikan. Sejauh ini, KAI menyatakan bahwa tidak ada bukti yang menunjukkan bahwa data pribadi penumpang atau karyawan telah terekspos. Untuk menyelidiki masalah ini, KAI bekerja sama dengan Badan Siber dan Sandi Negara (BSSN) dan unit kejahatan siber Kepolisian Negara Republik Indonesia atau Badan Reserse Kriminal (Bareskrim). Meskipun penyerang mengklaim telah mencuri data sensitif dan meminta tebusan sebesar 11,69 Bitcoin. Kasus permintaan tebusan menggunakan Bitcoin menunjukkan bahwa penjahat siber sering menggunakan mata uang kripto untuk mendapatkan keuntungan. Ini adalah metode umum yang dipilih oleh penyerang karena dianggap lebih efisien dan lebih aman bagi pihak mereka sendiri [50]. Namun, KAI tidak mudah mempercayai pihak yang membuat ancaman. Pernyataan resmi dari KAI menegaskan bahwa, hingga saat ini, tidak ada bukti yang menunjukkan bahwa ada data yang bocor seperti yang diklaim oleh tersangka penyerang. Namun, KAI tetap berkomitmen untuk melakukan

penyelidikan menyeluruh untuk mencari tahu kebenaran di balik masalah ini. KAI memastikan bahwa operasional TI-nya, termasuk layanan tiket daring dan sistem *Face Recognition Boarding Gate*, tetap berjalan dengan aman dan normal. KAI juga telah menerapkan *International Organization for Standardization* (ISO) yang merupakan standar internasional untuk sistem manajemen keamanan informasi atau yang lebih umum disebut sebagai *Information Security Management Systems* (ISMS) untuk melindungi data digitalnya.

Selain membocorkan dan mengancam, kata kerja lain yang menonjol adalah menyerang, menjual, mengakses, dan mencuri. Korpus tersebut mengungkapkan bahwa pengungkapan data tanpa izin juga mencakup isu-isu serangan siber, penjualan gelap, akses ilegal, dan pencurian data. Kata kerja seperti mengeksploitasi dan meretas juga terdapat dalam korpus tersebut, tetapi bukan termasuk istilah yang paling umum digunakan. Nama-nama individu dan lembaga ditampilkan secara mencolok dalam pelaporan korpus tentang pengungkapan data tanpa izin. Khususnya, nama-nama presiden Indonesia periode 2014–2024, Joko Widodo, dan menteri keuangannya, Sri Mulyani, muncul dengan frekuensi tinggi. Namun, korpus tersebut menunjukkan bahwa keduanya merupakan korban kebocoran data Nomor Pokok Wajib Pajak (NPWP) akibat kerentanan keamanan di lingkungan Direktorat Jenderal Pajak. Tokoh lain yang sering disebut adalah Bjorka, yang digambarkan sebagai aktor di balik kebocoran data tersebut. Bjorka mendapatkan popularitas sebagai sosok anonim yang mengaku memiliki data pejabat tinggi. Lembaga yang paling sering disebutkan dalam korpus tersebut antara lain BSSN dengan frekuensi tertinggi (596), Kementerian Komunikasi dan Informatika (Kominfo), Komisi Pemilihan Umum, Direktorat Jenderal Pajak, PT Kereta Api Indonesia, Pusat Data Nasional Sementara (PDNS), Pusat Data Nasional (PDN), dan Biznet. Selain BSSN, lembaga-lembaga ini terutama diidentifikasi sebagai korban pengungkapan data tanpa izin.

Kebocoran data yang melibatkan Nomor Pokok Wajib Pajak (NPWP) Presiden Joko Widodo dan Menteri Keuangan Sri Mulyani, yang diikuti oleh kebocoran data di Kementerian Komunikasi dan Informatika (Kominfo), berpotensi semakin menggerogoti kepercayaan publik terhadap kemampuan pemerintah dalam mengelola data secara efektif. Kekhawatiran ini diperparah oleh serangan siber yang menyasar lembaga-lembaga kunci lainnya, seperti Pusat Data Nasional Sementara (PDNS) dan Komisi Pemilihan Umum (KPU). Kata "publik" sendiri muncul sebanyak 52 kali di antara 212 artikel yang dikorpus. Korpus tersebut bahkan mencatat bahwa publik menuntut akuntabilitas dari aktor-aktor yang tidak secara eksplisit tercermin dalam analisis frekuensi kata, seperti Budie Arie, Menteri Komunikasi dan Informatika, yang didesak untuk mengundurkan diri. BSSN sering disebutkan dalam korpus tersebut, karena secara konsisten ditugaskan oleh Presiden untuk memitigasi dampak kebocoran data.

Gambar 3. Hasil *Word cloud*

Secara keseluruhan, kata kerja muncul lebih sering daripada nama lembaga, individu, atau kata benda lainnya, menjadikan kata kerja kategori dominan dalam analisis awan kata (Gambar 3 dan 4). Hal ini terlihat dari dominannya kata-kata seperti "*leaking*", yang muncul dalam ukuran font lebih besar daripada istilah lainnya. Hal ini menunjukkan bahwa korpus tidak hanya menekankan lembaga atau individu yang terlibat dalam insiden, tetapi juga memberikan perhatian signifikan pada tindakan dan respons para pelaku. Dengan demikian, kata kerja muncul sebagai elemen yang paling menonjol dan berulang dalam awan kata, menyoroti fokus pada bagaimana para pelaku

merespons pelanggaran tersebut. Hasil analisis *word cloud* ini telah lebih baik daripada beberapa penelitian sebelumnya. Hasil analisis *word cloud* Chandra dan Yusuf [51] masih memunculkan karakter tidak bermakna seperti "m" dan "?". Putri [52] masih memunculkan kata yang bersifat fungsional dan tidak bermakna substantif, seperti "di", "yang", "melalui", "atau", dan "ke". Hal ini karena proses pembersihan data teks (*preprocessing*) belum optimal, seperti tidak menghapus *stopword*, tanda baca, dan karakter tunggal, serta tokenisasi yang kurang tepat.

Gambar 4. Hasil *Word cloud*

Insiden pelanggaran data yang berulang juga mengungkapkan pola naratif dalam korpus: lembaga yang terdampak merespons dengan menekankan upaya mitigasi, para ahli merekomendasikan penguatan langkah-langkah keamanan siber, sementara pelaku mengeluarkan ancaman dan menuntut tebusan. Meskipun istilah teknis terkait aktivitas peretasan, seperti *ransomware*, *phishing*, dan *malware*, muncul dalam korpus, frekuensinya tidak memenuhi ambang batas harus lebih dari 49 yang telah ditetapkan dalam metodologi penelitian. Ambang batas ini diterapkan untuk mengecualikan istilah-istilah yang jarang digunakan dan dianggap memiliki nilai analitis terbatas. Meskipun demikian, istilah-istilah teknis tersebut berpotensi meningkatkan pemahaman publik tentang pelanggaran data dan akses tidak sah ke informasi.

5. KESIMPULAN

Dengan menganalisis 212 artikel dari empat media berita utama, istilah "kebocoran" muncul sebagai yang paling dominan, menunjukkan fokus utama media pada insiden kebocoran data. Kata kerja seperti "akses", "jual", dan "curi" muncul lebih sering daripada nama lembaga atau individu, menunjukkan bahwa narasi media lebih menekankan tindakan dan respons terhadap insiden daripada aktornya sendiri. Lembaga pemerintah dan tokoh masyarakat biasanya dibingkai sebagai korban, sementara aktor seperti Bjorka digambarkan sebagai pelaku. Temuan ini menunjukkan bahwa liputan media cenderung menyoroti sifat tindakan yang terlibat dalam pengungkapan data tanpa izin, terutama melalui penggunaan kata kerja tindakan yang dominan, daripada identitas korban atau pihak yang bertanggung jawab. Penelitian di masa mendatang didorong untuk mengkaji dinamika sentimen dan pembingkaian dalam liputan berita, serta membandingkan pola media di berbagai media atau periode waktu untuk mengamati pergeseran konstruksi wacana dari waktu ke waktu.

REFERENSI

[1] A. M. Admar, Sirojuzilam, Badaruddin, and Rujiman, "Study of the digitalization of population administration services using population data in Medan City regional development planning," *International Conference on Sciences Development and Technology*, vol. 3, no. 1, pp. 143–150, 2023, [Online]. Available: http://creativecommons.org/licenses/by-sa/4.0/

[2] Y. Gong, J. Yang, and X. Shi, "Towards a comprehensive understanding of digital transformation in government: Analysis of flexibility and enterprise architecture," *Gov Inf Q*, vol. 37, no. 3, Jul. 2020, doi: 10.1016/j.giq.2020.101487.

[3] C. Yang, M. Gu, and K. Albitar, "Government in the digital age: Exploring the impact of digital transformation on governmental efficiency," *Technol Forecast Soc Change*, vol. 208, Nov. 2024, doi: 10.1016/j.techfore.2024.123722.

[4] Y. Jiang *et al.*, "Pervasive user data collection from cyberspace: privacy concerns and countermeasures," *Cryptography*, vol. 8, no. 1, Mar. 2024, doi: 10.3390/cryptography8010005.

[5] S. N. I. Purnamaningsih, J. Ismono, I. S. Utami, V. Parlindungan, and S. N. Hanifah, "The challenges of data privacy laws in the age of big data: Balancing security, privacy, and innovation," *Join: Journal of Social Science*, vol. 1, no. 6, pp. 455–465, 2024, doi: https://doi.org/10.59613/gny8bq82.

[6] O. L. van Daalen, "The right to encryption: Privacy as preventing unlawful access," *Computer Law and Security Review*, vol. 49, Jul. 2023, doi: 10.1016/j.clsr.2023.105804.

[7] Surfshark, "Global data breach statistics," Amsterdam, 2025. [Online]. Available: https://surfshark.com/research/data-breach-monitoring?srsltid=AfmBOoqBjqTpUiD1_gtSj60xnVJfClo3BvkDKjIsJ-fvHz9wQILvd2pL

[8] C. Lubbers, Y. Lu, C. A. Lubbers, and A. Mcnamara, "A study of organizational responses to major data breaches in the retail & healthcare industries," *Quarterly Review of Business Disciplines*, vol. 3, no. 2, pp. 101–116, 2016, [Online]. Available: https://www.researchgate.net/publication/337290795

[9] H. R. Nikkhah and V. Grover, "Strategizing responses to data breaches: A multi-method study of organizational responsibility and effective communication with stakeholders," *Journal of Management Information Systems*, vol. 41, no. 4, pp. 1042–1077, 2024, doi: 10.1080/07421222.2024.2415774.

[10] V. Ibrahim, J. A. Bakar, N. H. Harun, and A. F. Abdulateef, "A word cloud model based on hate speech in an online social media environment," *Baghdad Science Journal*, vol. 18, pp. 937–946, Jun. 2021, doi: 10.21123/bsj.2021.18.2(Suppl.).0937.

[11] J. Jussila, A. H. Suominen, A. Partanen, and T. Honkanen, "Text analysis methods for misinformation-related research on Finnish language twitter," *Future Internet*, vol. 13, no. 6, Jun. 2021, doi: 10.3390/fi13060157.

[12] G. Sensales, G. Di Cicco, and C. Baldner, "Representations of Italian populism and immigration on Facebook. A comparison of the posts by Luigi Di Maio and Matteo Salvini (2014-2018)," *Papers on Social Representations*, vol. 30, no. 1, 2021, [Online]. Available: http://psr.iscte-iul.pt/index.php/PSR/index]2.1[http://psr.iscte-iul.pt/index.php/PSR/index]

[13] A. I. Kabir, K. Ahmed, and R. Karim, "Word cloud and sentiment analysis of amazon earphones reviews with R programming language," *Informatica Economică*, vol. 24, no. 4/2020, pp. 55–71, Dec. 2020, doi: 10.24818/issn14531305/24.4.2020.05.

[14] S. Wassan, X. Chen, T. Shen, M. Waqar, and N. Z. Jhanjhi, "Amazon product sentiment analysis using machine learning Techniques," *Revista Argentina de Clínica Psicológica*, vol. 30, no. 1, pp. 695–703, 2021, doi: 10.24205/03276716.2020.2065.

[15] Fadly, D. Marlina, T. B. Kurniawan, M. Z. Zakaria, and S. F. binti Abdullah, "Sentiment analysis on natural skincare products," *JOURNAL OF DATA SCIENCE*, vol. 2022, no. 12, pp. 2805–5160, 2022.

[16] N. M. Zafri, S. Afroj, I. M. Nafi, and M. M. U. Hasan, "A content analysis of newspaper coverage of COVID-19 pandemic for developing a pandemic management framework," *Heliyon*, vol. 7, no. 3, Mar. 2021, doi: 10.1016/j.heliyon.2021.e06544.

[17] F. Ahmed, M. Mubeen, and M. Nawaz, "Framing South Asian politics: An analysis of Indian and Pakistani English print media discourses regarding Kartarpur corridor," *PLoS One*, vol. 17, no. 2 February, Feb. 2022, doi: 10.1371/journal.pone.0264115.

[18] S. Mirza, C. Villa, and C. Pöpper, "Media talks privacy: Unraveling a decade of privacy discourse around the world," in *Proceedings on Privacy Enhancing Technologies*, Privacy Enhancing Technologies Symposium Advisory Board, Oct. 2024, pp. 102–122. doi: 10.56553/popets-2024-0109.

[19] L. Tavadze, I. Diasamidze, N. Katamadze, and L. Davitadze, "Modern tendencies in media discourse," *International Journal of Innovative Technologies in Social Science*, vol. 1, no. 41, 2024, doi: 10.31435/rsglobal_ijitss/30032024/8123.

[20] C. Reynolds, "Building theory from media ideology: Coding for power in journalistic discourse," *Journal of Communication Inquiry*, vol. 43, no. 1, pp. 47–69, Jan. 2019, doi: 10.1177/0196859918774797.

[21] M. Krzyżanowski and J. A. Tucker, "Re/constructing politics through social & online media: Discourses, ideologies, and mediated political practices," *Journal of Language and Politics*, vol. 17, no. 2, pp. 141–154, 2018.

[22] Y. Li and Q. Liu, "A comprehensive review study of cyber-attacks and cyber security; Emerging trends and recent developments," *Energy Reports*, vol. 7, pp. 8176–8186, Nov. 2021, doi: 10.1016/j.egyr.2021.08.126.

[23] B. Schneier, *Secrets and lies: Digital security in a networked world*, 1st ed. Wiley, 2004.

[24] T. Sasi, A. H. Lashkari, R. Lu, P. Xiong, and S. Iqbal, "A comprehensive survey on IoT attacks: Taxonomy, detection mechanisms and challenges," *Journal of Information and Intelligence*, Nov. 2023, doi: 10.1016/j.jiixd.2023.12.001.

[25] William. Stallings, *Network security essentials: Applications and standards*, Sixth. Pearson, 2017.

[26] K. Coulibaly, "An overview of intrusion detection and prevention systems," *arXiv preprint arXiv:2004.08967*, 2020, doi: https://doi.org/10.48550/arXiv.2004.08967.

[27] G. K. Zipf, *Human behavior and the principle of least effort*. Cambridge: Addison-Wesley Press, Inc, 1949.

[28] I. Moreno-Sánchez, F. Font-Clos, and Á. Corral, "Large-scale analysis of Zipf's law in English texts," *PLoS One*, vol. 11, no. 1, Jan. 2016, doi: 10.1371/journal.pone.0147073.

[29] A. Hossain, M. Karimuzzaman, M. M. Hossain, and A. Rahman, "Text mining and sentiment analysis of newspaper headlines," *Information (Switzerland)*, vol. 12, no. 10, Oct. 2021, doi: 10.3390/info12100414.

[30] M. L. Jockers and Thalken Rosamond, *Text analysis with R for students of literature*, Second. Springer, 2020. doi: https://doi.org/10.1007/978-3-030-39643-5.

[31] H. Heryono, D. Suganda, S. Yuliawati, and N. Darmayanti, "Word frequencies in linguistic articles published in SINTA indexed journals," *Jurnal Sosioteknologi*, vol. 22, no. 1, Apr. 2023, doi: 10.5614/sostek.itbj.2023.22.1.8.

[32] A. Roy, "Word frequency- content analysis approach to identify the terms in the content of library rules and regulation of Indian universities and European universities: A comparative study," *Library Philosophy and Practice (e-journal)*, pp. 1–15, 2023.

[33] B. Başaran, "Measuring word frequency in language teaching textbooks using LexiTürk," *International Online Journal of Education and Teaching (IOJET)*, vol. 9, no. 1, pp. 2148–225, 2022.

[34] R. M. M. Hicke, M. Goenka, and E. Alexander, "Word clouds in the wild," in *2022 IEEE 7th Workshop on Visualization for the Digital Humanities (VIS4DH)*, Oct. 2022, pp. 43–48. doi: https://doi.org/10.1109/VIS4DH57440.2022.00015.

[35] F. Guneri, G. Hayırcıl, M. U. Deligny, and S. Iman, "A word visualization observation: The hidden meaning of the words in work environments," *European Scientific Journal ESJ*, vol. 17, no. 36, Oct. 2021, doi: 10.19044/esj.2021.v17n36p19.

[36] D. Nguyen *et al.*, "How we do things with words: Analyzing text as social and cultural data," *Front Artif Intell*, vol. 3, no. 62, Aug. 2020, doi: 10.3389/frai.2020.00062.

[37] T. Dingler, D. Kern, K. Angerbauer, and A. Schmidt, "Text priming-effects of text visualizations on readers prior to reading," in *16th IFIP Conference on Human-Computer Interaction (INTERACT)*, 2017, pp. 345–365. doi: https://doi.org/10.1007/978-3-319-67687-6_23.

[38] D. Knight, N. Khallaf, P. Rayson, M. El-Haj, I. Ezeani, and S. Morris, "FreeTxt: A corpus-based bilingual free-text survey and questionnaire data analysis toolkit," *Applied Corpus Linguistics*, vol. 4, no. 3, Dec. 2024, doi: 10.1016/j.acorp.2024.100103.

[39] A. H. Sasongko, R. Syarief, N. T. Rochman, and S. Amanah, "Word cloud analysis and business model of science techno park in Indonesian universities," *International Journal of Business Studies*, vol. 9, no. 2, pp. 137–148, 2025.

[40] S. Rahmadani and I. Yuadi, "Understanding political narratives: Word cloud analysis of Yoon Seok-Yeol's impeachment," *Journal of Law, Politic and Humanities*, vol. 5, no. 4, pp. 2687–2695, 2025, doi: 10.38035/jlph.v5i4.

[41] B. B. Sağlam, "Word cloud analysis of special issues in maritime transportation journals," *Journal of Transportation and Logistics*, vol. 10, no. 1, pp. 211–221, 2025, doi: 10.26650/JTL.2025.

[42] R Core Team, "R: A language and environment for statistical computing," 2024, 26.2.4. [Online]. Available: https://www.R-project.org/

[43] J. Silge and D. Robinson, "tidytext: Text mining and analysis using tidy data principles in R," *The Journal of Open Source Software*, vol. 1, no. 3, p. 37, Jul. 2016, doi: 10.21105/joss.00037.

[44] H. Wickham, R. François, L. Henry, K. Müller, and D. Vaughan, "dplyr: A Grammar of data manipulation," 2023, 1.1.4. [Online]. Available: https://dplyr.tidyverse.org

[45] Irvandi, B. Irawan, and O. Nurdiawan, "Naive bayes and wordcloud for sentiment analysis of halal tourism in Lombok Island Indonesia," *Innovation in Research of Informatics (INNOVATICS)*, vol. 5, no. 1, pp. 30–35, 2023.

[46] J. A. Wibowo, V. C. Mawardi, and T. Sutrisno, "Visualisasi word cloud hasil analisis sentimen berbasis fitur layanan aplikasi gojek dengan support vector machine," *Jurnal Serina Sains, Teknik dan Kedokteran*, vol. 2, no. 1, pp. 61–70, Mar. 2024, doi: 10.24912/jsstk.v2i1.32058.

[47] M. Fathoni and Syarifuddin, "Word cloud analysis for the urgency of the role of islamic schools headmaster in Merdeka curriculum implementation," *Jurnal 12 Waiheru*, vol. 10, no. 1, pp. 57–72, 2024.

[48] D. Lang and G. Chien, "wordcloud2: Create word cloud by 'htmlwidget,'" 2018, 0.2.1. [Online]. Available: https://github.com/lchiffon/wordcloud2

[49] S. Nazah, S. Huda, J. Abawajy, and M. M. Hassan, "Evolution of dark web threat analysis and detection: A systematic approach," *IEEE Access*, vol. 8, pp. 171796–171819, 2020, doi: 10.1109/ACCESS.2020.3024198.

[50] N. Katagiri, "From prepaid cards to bitcoin: How did ransomware hackers adopt cryptocurrencies?," *Journal of Cyber Policy*, vol. 9, no. 2, pp. 239–255, May 2024, doi: 10.1080/23738871.2024.2435956.

[51] M. I. A. Chandra and R. Yusuf, "Visualization of keywords in the 2024 election news using spacy and wordcloud," *TEKNIMEDIA*, vol. 5, no. 1, pp. 41–46, 2024.

[52] N. A. R. Putri, "Identifikasi topik dominan pada hastag konten anak di media sosial X menggunakan word cloud," *Indonesian Journal Of Information Technology*, vol. 3, no. 1, 2025.