Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 960−971
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7473

Analisis Sentimen Masyarakat Terhadap Kebocoran Pusat Data Nasional Sementara Menggunakan Algoritma Random Forest dan Support Vector Machine

Faishal Khairi Basri*, M Afdal, Angraini Angraini, Nesdi Evrilyan Rozanda

Fakultas Sains dan Teknologi, Sistem Informasi, Universitas Islam Negeri Sultan Syarif Kasim, Riau, Indonesia
Email: [1,*]khairibasri302@gmail.com, [2]m.afdal@uin-suska.ac.id, [3]angraini@uin-suska.ac.id, [4]nesdi.er@uin-suska.ac.id
Email Penulis Korespondensi: khairibasri302@gmail.com
Submitted: **31/05/2025**; Accepted: **01/09/2025**; Published**: 02/09/2025**

Abstrak−Serangan *ransomware* terhadap Pusat Data Nasional Sementara (PDNS) pada Juni 2024 menimbulkan kekhawatiran besar di masyarakat terkait keamanan data dan kesiapan pemerintah dalam menghadapi ancaman *cyber*. Penelitian ini bertujuan untuk menganalisis sentimen publik terhadap insiden tersebut menggunakan pendekatan analisis sentimen berbasis aspek pada 2.700 *tweet* berbahasa Indonesia yang dikumpulkan dari platform X. Penelitian mengikuti tahapan SEMMA (*Sample, Explore, Modify, Model, Assess*) yang mencakup pra-pemrosesan teks, ekstraksi aspek menggunakan *Part-of-Speech Tagging* dan pengenalan entitas bernama, representasi fitur menggunakan *Term Frequency-Inverse Document Frequency*, serta penyaringan aspek berdasarkan koherensi semantik. Aspek yang diperoleh dikelompokkan ke dalam lima kategori: keamanan data, tokoh dan lembaga, infrastruktur, politik dan ekonomi, serta dampak. Klasifikasi sentimen dilakukan menggunakan model IndoBERTweet. Hasil analisis menunjukkan dominasi sentimen negatif, khususnya pada aspek infrastruktur dan tokoh-lembaga, serta tidak ditemukannya sentimen positif pada aspek politik dan ekonomi. Untuk mengatasi ketidakseimbangan distribusi sentimen, digunakan metode *Synthetic Minority Oversampling Technique* dalam tahap pelatihan model. Evaluasi terhadap dua algoritma, yaitu *Random Forest* dan *Support Vector Machine*, menunjukkan bahwa *Random Forest* memiliki performa terbaik dengan akurasi 96% pada rasio data 70:30 dan rata-rata akurasi 99,05% pada validasi silang *10-fold*. Temuan ini menegaskan efektivitas analisis sentimen berbasis aspek dan keunggulan Random Forest dalam mengklasifikasikan data yang kompleks dan tidak seimbang.

Kata Kunci: Analisis Sentimen; Analisis Sentimen Berbasis Aspek; Kebocoran Data; *Random Forest*; *Support Vector Machine*

Abstract−A ransomware attack on Indonesia's Temporary National Data Center (PDNS) in June 2024 triggered major public concern over data security and government preparedness. This study aims to analyze public sentiment toward the incident using an Aspect-Based Sentiment Analysis approach on 2,700 Indonesian-language tweets collected from the X platform. The research follows the SEMMA (Sample, Explore, Modify, Model, Assess) methodology, involving text preprocessing, aspect extraction using part-of-speech tagging and named entity recognition, feature representation using Term Frequency-Inverse Document Frequency, and aspect refinement through semantic coherence. Extracted aspects are grouped into five categories: data security, institutions, infrastructure, politics and economy, and impact. Sentiment classification is carried out using the IndoBERTweet model. Results indicate a strong dominance of negative sentiment, particularly in the infrastructure and institutional categories, with no positive sentiment recorded in the political and economic aspect. To address class imbalance in sentiment distribution, the Synthetic Minority Oversampling Technique is applied during model training. Performance evaluation of two algorithms—Random Forest and Support Vector Machine—shows that Random Forest performs best, achieving 96% accuracy on a 70:30 data split and 99.05% average accuracy using 10-fold cross-validation. These findings highlight the effectiveness of aspect-based sentiment analysis and demonstrate Random Forest's superiority in handling imbalanced sentiment classification tasks.

Keywords: Aspect-Based Sentiment Analysis; Data Breach; Random Forest; Sentiment Analysis; Support Vector Machine

1. PENDAHULUAN

Zaman modern telah membawa perkembangan teknologi yang sangat pesat, termasuk dominasi media sosial dalam interaksi manusia [1]. Saat menggunakan internet, keamanan data menjadi aspek penting yang perlu diperhatikan, khususnya ketika berurusan dengan data-data pribadi yang bersifat sensitif. Kelalaian akan keamanan data pribadi dapat menimbulkan berbagai macam kerugian dengan skala yang beragam [2]. Salah satu insiden besar terjadi pada 20 Juni 2024, ketika Pusat Data Nasional Sementara (PDNS) mengalami serangan *ransomware* oleh kelompok Brain Chipper yang menuntut tebusan 8 juta untuk membuka data yang dikuncinya.

Kebocoran data bukan hal baru di Indonesia. Sejak tahun 2020, berbagai kasus besar telah terjadi, seperti kebocoran 91 juta data pengguna Tokopedia dan 279 juta data BPJS Kesehatan [3]. Kejadian-kejadian ini memengaruhi kepercayaan publik terhadap tata kelola data pemerintah. Kepercayaan publik merupakan faktor penting dalam menghasilkan legitimasi, yang dapat menghasilkan aset sosial untuk pemerintah dalam politik dan kegiatan pemerintah. Dalam hal ini, analisis sentimen diperlukan untuk mengetahui tingkat kepercayaan masyarakat yang ada [4]. Analisis sentimen menjadi alat penting dalam memahami opini publik di media sosial. Dengan mengelompokkan opini menjadi sentimen positif, negatif, atau netral, pendekatan ini telah digunakan dalam berbagai bidang seperti pemasaran, politik, kebijakan publik dan telah beralih ke teks media sosial, salah satunya platform X [5][6]. Namun demikian, klasifikasi sentimen yang bersifat umum belum cukup dalam mengungkap letak subtansial persoalan, terutama dalam kasus yang melibatkan berbagai dimensi tematik seperti insiden kebocoran data nasional. Oleh karena itu, pendekatan *Aspect-Based Sentiment Analysis* (ABSA) menjadi relevan karena memungkinkan analisis sentimen yang lebih rinci terhadap aspek tertentu dalam teks [7].

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 960−971
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7473

Platform X telah menjadi platform yang ideal untuk mencermati reaksi masyarakat terhadap kebijakan proteksi data pribadi, karena mereka memungkinkan analisis sentimen yang luas dan *real-time* [8]. Pengguna X sering berbagi pendapat, pemikiran, dan tanggapan mereka secara terbuka, yang dapat digunakan untuk mempelajari perasaan mereka tentang topik tertentu atau komentar mereka tentang layanan tertentu [9]. Dengan menggunakan analisis sentimen pada X, penelitian ini dapat mengevaluasi efektivitas strategi proteksi data yang saat ini digunakan dan memberikan rekomendasi untuk meningkatkan kerja sama antara pemerintah, institusi, dan masyarakat dalam melindungi data pribadi [10].

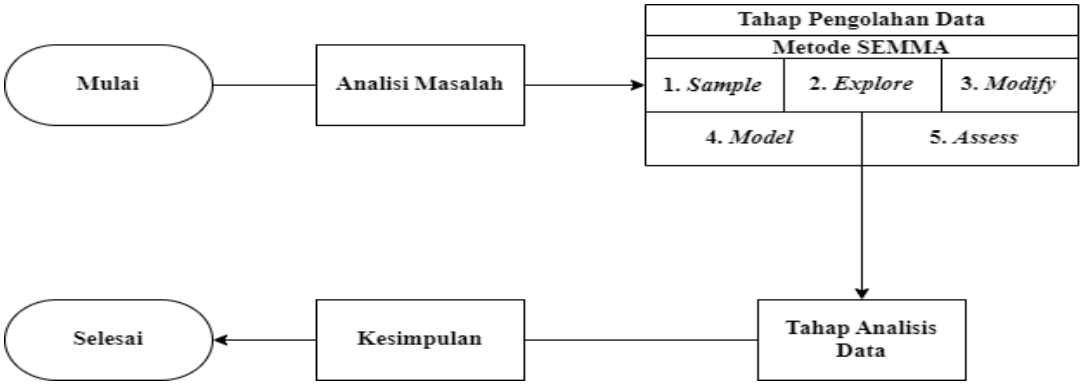
Beberapa penelitian sebelumnya telah menggunakan analisis sentimen dalam konteks kebocoran data. Seperti penelitian yang dilakukan oleh Sholehurrohman dan Ilman (2022) yang menganalisis kasus kebocoran data pengguna Facebook dengan menggunakan *web scraping* API dan klasifikasi sentimen [11]. Selain itu, Ahmad Turmudi Zy dan Wahyu Hadikristanto (2023) meneliti persepsi publik terhadap isu kebocoran data menggunakan algoritma *Support Vector Machine* (SVM) dan *Naïve Bayes*, dengan hasil presisi masing-masing sebesar 80% dan 97% [12]. Penelitian lain oleh Umam et al. (2024) membandingkan performa *Random Forest* dan *Support Vector Classification* (SVC) dalam klasifikasi *email phishing*, di mana SVC konsisten menunjukkan akurasi tertinggi hingga 97,52% [13]. Dalam konteks lokal, Alvali Zaqi Taufan dan Wahyu Wibowo (2024) mengkaji sentimen masyarakat terhadap keamanan data dan privasi pasca disahkannya UU PDP. Penelitian ini menemukan bahwa Linear SVM menjadi model paling akurat dengan akurasi 92,6% dan AUC 97% [14]. Penelitian terkait lainnya oleh Ichlasul Amal et al. (2022) juga menunjukkan bahwa SVM unggul dalam menganalisis sentimen terhadap isu kebocoran data kartu SIM, dengan *F1-score* tertinggi dibanding *Random Forest*, *Logistic Regression*, dan IndoBERT [15].

Meskipun sejumlah penelitian terdahulu telah menerapkan analisis sentimen, sebagian besar masih terbatas pada level dokumen atau kalimat, tanpa mengidentifikasi aspek-aspek spesifik yang menjadi perhatian publik. Selain itu, model klasik yang digunakan umumnya belum mempertimbangkan ketidakseimbangan distribusi kelas sentimen, sehingga cenderung bias terhadap kelas mayoritas. Studi ini mengadopsi pendekatan ABSA yang dikembangkan untuk konteks Indonesia, dengan mengintegrasikan ekstraksi aspek berbasis *Part Of Speech Tagging* dan *Named Entity Recognition* (NER), kemudian diperkuat dengan pembobotan menggunakan TF-IDF serta evaluasi dengan *semantic coherence* untuk memastikan keterhubungan antar aspek. Proses klasifikasi sentimen dilakukan dengan model *pretrained* IndoBERTweet, yang dikembangkan secara khusus untuk menangani dinamika bahasa di media sosial berbahasa Indonesia. Untuk menilai kualitas hasil klasifikasi sekaligus menjaga konsistensi performa model, digunakan dua algoritma pembelajaran mesin, *Random Forest* dan *Support Vector Machine* (SVM), yang telah terbukti andal dalam mengelola data teks berukuran menengah dan bersifat *nonlinier*. Selain itu, teknik *Synthetic Minority Oversampling Technique* (SMOTE) diterapkan agar distribusi sentimen yang tidak seimbang tetap terwakili secara proporsional selama proses pelatihan.

Dengan menggunakan data dari platform X dalam periode 20 Juli–31 Desember 2024, penelitian ini bertujuan untuk menganalisis respons publik terhadap insiden kebocoran PDNS secara terstruktur. Hasil analisis diharapkan dapat memberikan masukan untuk pengembangan strategi perlindungan data yang lebih efektif dan responsif terhadap kebutuhan masyarakat [16]. Selain memetakan sentimen berdasarkan aspek, analisis juga diarahkan untuk mengungkap ekspresi emosional yang muncul, seperti kemarahan, kekecewaan, dan ketidakpercayaan, yang secara konsisten tampak dalam sentimen negatif terhadap aspek-aspek tertentu. Melalui pendekatan ini, diharapkan diperoleh pemahaman yang lebih menyeluruh mengenai persepsi publik terhadap tata kelola keamanan siber nasional, sekaligus memberikan kontribusi konseptual dalam pengembangan metode analisis sentimen berbasis aspek dalam konteks kebijakan digital di Indonesia.

2. METODOLOGI PENELITIAN

Metodologi penelitian merupakan alur-alur penilitian yang menggambarkan langkah-langkah dalam melaksanakan penelitian dari awal hingga selesai agar penelitian dapat dilaksanakan secara sistematis dan mencapai tujuan. Metodologi penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Alur Penelitian

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 960−971
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7473

2.1 Analisis Masalah

Tahap awal penelitian ini difokuskan pada pemahaman konteks dan identifikasi akar permasalahan yang berkaitan dengan insiden kebocoran Pusat Data Nasional Sementara, sebagai dasar dalam merumuskan tujuan, ruang lingkup, dan urgensi penelitian. Setelah itu, dilakukan Studi Literatur yang bertujuan untuk mengidentifikasi fakta-fakta penting dan kesenjangan dalam penelitian sebelumnya yang dapat menjadi latar belakang dan alasan pendukung penelitian.

2.2 Tahap Pengolahan Data (Metode SEMMA)

Metode SEMMA juga dapat diartikan sebagai implementasi praktis dari lima tahap dalam proses KDD (*Knowledge Discovery in Database*) yang merupakan singkatan dari *sample*, *explore*, *modify*, *model*, dan *assess* [17]. Tahapan-tahapan ini memungkinkan untuk mengimplementasikan model yang kuat dan lebih akurat untuk pengambilan keputusan berdasarkan hasil yang diperoleh. Oleh karena itu, proses ini menentukan struktur yang teratur dan sistematis yang menangani proyek-proyek kompleks berdasarkan penggalian data [18].

a. *Sample*
   Pada tahapan *sample*, dilakukan pengumpulan data mengenai topik kebocoran PDNS dengan *web crawling* dari platform X. *Web Crawling* merupakan proses untuk mengumpulkan informasi yang ditampilkan di browser web secara otomatis yang bertujuan untuk mengekstrak informasi dari web [19], [20].
b. *Explore*
   Tahap ini bertujuan untuk mendiskripsikan data yang didapatkan melalui *web crawling* dari platform X tentang kebocoran Pusat Data Nasional Sementara dan menentukan data yang digunakan.
c. *Modify*
   Dalam tahapan *modify*, dilakukan *text preprocessing* yang di mulai dari tahap *case folding*, yaitu perubahan semua huruf dalam teks menjadi huruf kecil. Selanjutnya tahapan *cleaning*, yaitu menghapus data yang tidak relevan dan tidak memiliki makna dalam analisis teks. Lalu dilakukan *tokineze* untuk memecah teks menjadi token-token kata. Selanjutnya *stopword removal*, yaitu menghilangkan kata-kata umum yang tidak memberikan nilai tambahan pada analisis teks. Terakhir adalah *stemming*, yaitu tahapan untuk menghilangkan imbuhan atau awalan dari kata agar hanya tesrsisa akar kata (*root word*) atau bentuk dasarnya [21].
d. *Model*
   Tahap ini bertujuan untuk melakukan *Aspect-Based Sentiment Analysis* (ABSA). Ekstraksi aspek dilakukan menggunakan *Part of Speech Tagging* (POS Tag) dan *Named Entity Recognition* (NER), yang hasilnya disaring menggunakan pembobotan TF-IDF dan pengukuran *Semantic Coherence* untuk menjamin relevansi semantik. Aspek-aspek yang telah tersaring kemudian dikelompokkan secara manual ke dalam lima kategori tematik, yaitu Keamanan Data, Lembaga Pemerintah, Infrastruktur, Politik, serta Dampak. Kategori ini dipilih berdasarkan penelaahan dari beberapa penelitian seperti, Lee et al., (2022), Correa et al., (2025), Shandler et al., (2023), Prince et al., (2024) [22], [23], [24], [25].
   Setelah kategorisasi, pelabelan sentimen dilakukan secara otomatis menggunakan model *pre-trained* IndoBERTweet (Aardiiiiy/indobertweet-base-Indonesian-sentiment-analysis) yang telah disesuaikan untuk teks media sosial berbahasa Indonesia. Selanjutnya, dilakukan pembagian data latih dan data uji yang berfungsi untuk mengevaluasi kemampuan generalisasi model terhadap data. Melalui perbandingan performa pada berbagai rasio pembagian, analisis sensitivitas model terhadap ukuran data pelatihan juga dapat dilakukan sebagai bagian dari pengujian keandalan algoritma yang digunakan [26].
e. Assess
   Pada tahap *assess*, dilakukan penilaian kinerja performa terhadap algoritma *Random Forest* dan *Support Vector Machine* menggunakan beberapa matriks seperti akurasi, presisi, *recall*, dan *f1-score*. Setelah itu dilakukannya *K-Folds Cross Validation* untuk melakukan validasi performa kinerja dari algoritma yang dilakukan pada dataset.

2.3 Tahap Analisis Data

Tahap ini penulis merupakan tahap menganilisis dari hasil penelitian. Langkah pertama adalah analisa klasfikasi dengan menganilisis kembali hasil klasifikasi berdasarkan alur dari tahap-tahap penelitian. Setelah itu dilakukannya visualisasi klasifikasi menggunakan *Wordcloud. Wordcloud* merupakan visualisasi komputer yang menampilkan kata kunci utama dari teks atau dokumen pada kanvas visualisasi [27].

2.4 Kesimpulan

Kesimpulan disusun berdasarkan keseluruhan proses metodologis dan temuan penelitian, yang merefleksikan hubungan antara pendekatan yang diterapkan dan hasil yang diperoleh.

2.5 IndoBERTweet

Dalam beberapa tahun terakhir, ditemukannya metode *Bidirectional Encoder Representations form Transformers* (BERT) dalam bidang NLP yang melampaui metode tradisional. BERT meningkatkan pemahaman bahasa dan konteks, yang memungkinkan penggunaan model *pre-trained* yang lebih cepat dan mudah di aplikasikan dalam analisis sentimen [28]. Salah satu model BERT yang cocok dari *Hugging Face* untuk dataset dari platform X dan

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 960−971
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7473

berbahasa Indonesia adalah IndoBERTweet [29]. Dengan memanfaatkan *pipeline* dari *transformers*, setiap teks dianalisis untuk menentukan polaritas sentimennya terhadap aspek yang telah diidentifikasi sebelumnya [30].

2.6 POS Tag dan NER

Part of Speech Tagging (POS Tag) adalah teknik untuk mengelompokkan kata berdasarkan kelas kata seperti kata benda, kata kerja, atau kata sambung. POS Tag merupakan sistem berbasis pembelajaran mesin yang diciptakan untuk mengotomatisasi proses menandai aspek, yang dapat memakan waktu yang lama jika dilakukan secara manual [31]. Sedangkan *Named Entity Rcognition* (NER) berfokus pada mengekstrak informasi penting dari teks yang tidak terstruktur. NER bekerja untuk mengidentifikasi dan mengklasifikasikan komponen penting, seperti nama orang, lokasi geografis, tanggal, organisasi, dan peristiwa [32].

2.7 *Semantice Coherence*

Semantic Coherence merupakan teknik dalam NLP yang bertugas untuk penyaringan dan pemilihan informasi tertentu. *Semantic Coherence* menekankan bahwa kalimat harus koheren dan konsisten dengan mempertimbangkan elemen seperti topik, urutan kata, penggunaan bahasa yang tepat, konteks keseluruhan, dan struktur kalimat [7].

2.8 Pembobotan TF-IDF

Teknik TF-IDF (*Term Frequency-Inverse Document Frequency*) adalah teknik yang digunakan untuk menghitung bobot *term* dalam sebuah dokumen. Komponen TF berfungsi untuk menghitung frekuensi sebuah *term* dalam sebuah dokumen. Sedangkan komponen IDF berfungsi untuk menentukan signifikansi sebuah *term* dengan mempertimbangkan kemunculannya di dokumen dan membedakannya dari *stopwords*. Ini dihitung dengan mengambil logaritma dari rasio antara jumlah total dokumen dan jumlah dokumen yang mengandung *term*, seperti yang ditunjukkan dalam persamaan:

$$TF - IDF(t) = TF(t,d) \times IDF(t) \tag{1}$$

2.9 *Random Forest*

Random Forest adalah algoritma yang menggunakan metode *bagging* untuk mengambil beberapa sampel data, melatih beberapa pohon keputusan, dan melakukan pengambilan suara terbanyak untuk klasifikasi, sedangkan pada masalah regresi dilakukan pengambilan nilai rata-rata (*mean*) [33]. Dalam kasus masalah regresi, ketika menggunakan algoritma *Random Forest*, *Mean Squared Error* (MSE) digunakan untuk menentukan bagaimana data bercabang dari setiap *node* [34]. Rumus *Random Forest* digambarkan menjadi:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2 \tag{2}$$

2.10 *Support Vector Machine*

Support Vector Machines (SVM) merupakan salah satu algoritma untuk klasifikasi dan regresi yang telah diterapkan di banyak aplikasi berkat banyak fitur menarik dan kinerja empiris yang tinggi [35]. SVM menemukan batas yang memisahkan kelas-kelas yang berbeda satu sama lain di mana dalam ruang 2 dimensi, batas tersebut dinamai garis, dalam ruang 3 dimensi batas tersebut dinamai bidang, dan akhirnya dalam dimensi yang lebih besar dari 3 batas tersebut dinamai *Hyper Field* [34]. Rumus SVM digambarkan di bawah ini:

$$MINIMIZE_{a_{0,\dots}\,a_m} = \sum_{j=1}^{n} MAX\{0,1 - (\sum_{i=1}^{m} a_i x_{ij} + a_0) y_j\} + \lambda \sum_{i=1}^{m}(a_i)^2 \tag{3}$$

2.11 Metode SMOTE

Pada penelitian Vaishali dan Ratnavel (2024), seringkali mendapatkan penelitian ABSA yang memiliki distribusi polaritas sampel dalam dataset yang tidak seimbang [36]. Hal ini menyebabkan algoritma klasifikasi cenderung bias terhadap kelas mayoritas dan sulit untuk mengidentifikasi kelas minoritas. Oleh karena itu, perlu dilakukannya tindakan *resampling data* dalam menyeimbangkan distribusi kelas untuk mengatasi masalah ini [37]. *Resampling data* yang digunakan dalam penelitian ini adalah SMOTE yang menggunakan *Random Oversampling* untuk menghasilkan data baru berdasarkan pada kesamaan antara sejumlah kecil data.

2.12 *K-Fold Cross Validation*

K-Fold Cross Validation adalah teknik evaluasi model yang digunakan untuk menilai kinerja prediktif model [38]. Teknik ini membagi dataset menjadi k bagian (*fold*), yang masing-masing secara bergantian digunakan sebagai data uji, dan sisa *fold* lainnya digunakan untuk melatih model. Tujuan utama dari *K-Fold Cross Validation* adalah memberikan estimasi kinerja model yang lebih akurat dengan meminimalkan bias yang muncul akibat pembagian data yang tidak representative [39].

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 960−971
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7473

3. HASIL DAN PEMBAHASAN

3.1 Tahap Pengolahan Data (Metode SEMMA)

3.1.1 *Sample*

Dalam penelitian ini, data yang digunakan diperoleh dari tweet dengan kata kunci "Pusat Data Nasional" dan berbahasa Indonesia dalam rentang waktu 20 Juni – 31 Desember 2024. Jumlah total data yang diperoleh selama periode yang ditetapkan mencapai 2700 tweet.

3.1.2 *Explore*

Pada tahap sebelumnya, didapatkan 14 kolom dari *crawling* yang berisi *conversation_id_str*, *created_at*, *favorite_count*, *full_text*, *id_str*, *image_url*, *in_reply_to_screen_name*, *lang*, *location*, *quote_count*, *reply_count*, *retweet_count*, *tweet_url*, *user_id_str*, *username*. Penulis hanya menggunakan kolom *full_text* untuk di proses pada tahap selanjutnya. Hasil filter kolom ini dapat dilihat pada Tabel 1.

Tabel 1. Hasil Filter *Crawling Data*

No	full_text
1	Pembuat ransomware Brain Cipher menjanjikan akan membebaskan data dari Pusat Data Nasional Sementara (PDNS) 2 yang mereka sandera dari pemerintah. https://t.co/AfOOc4MIbo
2	GANGGUAN pada Pusat Data Nasional Sementara (PDNS) 2 di Surabaya berdampak terhambatnya proses sertifikasi halal pelaku usaha mikro kecil. https://t.co/S9xcZupXyK
3	Diretasnya Pusat Data Nasional Sementara (PDNS) ternyata berimbas pada terhambatnya tender (lelang) proyek pembangunan Ibu Kota Nusantara (IKN) Provinsi Kalimantan Timur (Kaltim). Baca di https://t.co/ufhQWHgP3s https://t.co/Ju7L2mRLtt
...	...
2700	cloud rusak, perlindungan mati, sistem kacau. Ini negara atau apa? 😭

3.1.3 *Modify*

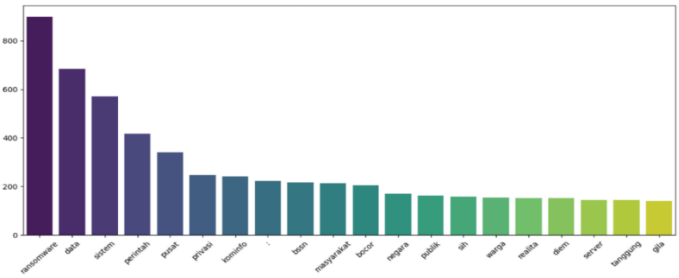
Pada tahap ini dilakukan modifikasi berupa *text preprocessing* pada dataset dengan tujuan agar dataset menjadi terstruktur. Hasil *text preprocessing* dapat dilihat pada Tabel 2.

Tabel 2. Hasil *Preprocessing Dataset*

full_text	Pembuat ransomware Brain Cipher menjanjikan akan membebaskan data dari Pusat Data Nasional Sementara (PDNS) 2 yang mereka sandera dari pemerintah. https://t.co/AfOOc4MIbo
Case Folding	pembuat ransomware brain cipher menjanjikan akan membebaskan data dari pusat data nasional sementara (pdns) 2 yang mereka sandera dari pemerintah. https://t.co/afooc4mibo
Cleaning	pembuat ransomware brain cipher menjanjikan akan membebaskan data dari pusat data nasional sementara pdns yang mereka sandera dari pemerintah
Tokenize	['pembuat', 'ransomware', 'brain', 'cipher', 'menjanjikan', 'akan', 'membebaskan', 'data', 'dari', 'pusat', 'data', 'nasional', 'sementara', 'pdns', 'yang', 'mereka', 'sandera', 'dari', 'pemerintah']
Stopword Removal	['pembuat', 'ransomware', 'brain', 'cipher', 'menjanjikan', 'membebaskan', 'data', 'pusat', 'data', 'nasional', 'pdns', 'sandera', 'pemerintah']
Stemming	['buat', 'ransomware', 'brain', 'cipher', 'janji', 'bebas', 'data', 'pusat', 'data', 'nasional', 'pdns', 'sandera', 'perintah']

3.1.4 *Model*

a. POS Tag dan NER
Setelah melalui tahap *text preprocessing*, digunakan *library stanza* berbahasa indonesia untuk melakukan POS Tag dan tranformasi IndoBERTweet untuk melakukan NER. Hasil 20 aspek teratas yang didapatkan dapat dilihat pada Gambar 2.



Gambar 2. Visualisasi 20 Aspek Teratas dari POS Tag dan NER

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 960−971
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7473

b. Pembobotan TF-IDF
Proses filter TF-IDF menghasilkan kata-kata sebagai calon aspek atau kandidat aspek berdasarkan bobot aspek yang didapatkan dari POS Tag dan NER. TF-IDF melakukan filter pada 208 aspek yang didapatkan sebelumnya dan mengurutkan dari bobot paling besar hingga bobot paling kecil. Berikut hasil 20 aspek tertinggi setelah filter TF-IDF: sistem, perintah, data, ransomware, privasi, masyarakat, bssn, kominfo, bocor, pusat, informasi, malware, warga, server, instansi, akses, nasional, sementara, cyber, negara. Distribusi pembobotan TF-IDF pada aspek-aspek yang didapatkan disajikan pada Tabel 3.

Tabel 3. Hasil Filter Pembobotan TF-IDF

No.	sistem	perintah	data	ransomware	...	negara
1	0	0	0	0,085097177	...	0,153773872
2	0	0	0	0,081471979	...	0,147223
3	0	0	0	0,092493784	...	0,167139825
...	...	...	...	...	...	...
2700	0,42098693	0	0	0	...	0,607234043

c. *Semantic Coherence*
Aspek-aspek yang telah diekstraksi, disaring berdasarkan nilai kemiripan semantik menggunakan *cosine similarity* antar vektor TF-IDF. Aspek yang didapatkan meliputi :agusyudhoyono, ahy, akses, akibat, ancam, antivirus, arie, baharu, backup, bpn, budi, bukti, cipher, *cyber*, daftar, dampak, s, demokrat, deteksi, dpr, enkripsi, ganggu, hacker, hutang, ibukota, ikn, informatika, insiden, internet, isi, it, jabat, kait, kala, kena, klaim, kelompok, kominfo, komisi, kunci, laku, layan, leaks, lindung, *lockbit*, lumpuh, maaf, *malware*, menteri, menkominfo, mulyono, nama, nasional, negara, pangerapan, pdemokrat, pdn, pdns, perintah, perangkat, politik, ppn, publik, pusat, qilin, rakyat, *ransomware*, retas, revolusi, ri, semuel, serang, server, siber, sistem, tanggung, tokoh, trojan, undur, upaya, usaha, virus, wapres.

d. Pengkategorian Aspek
Langkah selanjutnya adalah mengelompokkan data menjadi 5 kategori aspek. Langkah ini dilakukan dengan cara menyusun kata-kata kunci ke dalam kategori tematik yang lebih tinggi. Pada tahap ini, penulis secara eksplisit menentukan hubungan antar aspek berdasarkan pengetahuan domain dan konteks isu yang dibahas. Hasil pengkategorian secara lengkap dapat dilihat pada Tabel 4.

Tabel 4. Pengkategorian Aspek

Aspek	Kata Tiap Aspek	Kategori
1	ransomware, malware, trojan, hacker, retas, serang, virus, cyber, siber, insiden, cipher, enkripsi, deteksi, antivirus, backup, lindung, leaks, lockbit, qilin	Keamanan Data
2	menteri, menkominfo, wapres, dpr, komisi, ri, nasional, negara, kominfo	Lembaga
3	sistem, server, data, informatika, internet, perangkat, akses, isi, pusat, pdn, pdns, it, kunci, layan, upaya	Infrastruktur
4	bpn, rakyat, daftar, publik, jabat, perintah, ikn, ibukota, usaha, dpr, kelompok	Politik
5	dampak, akibat, ganggu, lumpuh, ancam, undur, maaf, usaha, tanggung, klaim, bukti, kena, revolusi, laku, kait, kala	Dampak

e. Pelabelan Data Berdasarkan Kategori Aspek
Distribusi hasil pelabelan sentimen menunjukkan dominasi opini negatif pada hampir seluruh aspek. Aspek Infrastruktur dan Tokoh dan Lembaga Pemerintah mencatat jumlah tertinggi dengan masing-masing 1.111 dan 1.054 *tweet* negatif. Aspek Politik dan Ekonomi juga menunjukkan kecenderungan negatif yang kuat, tanpa ada sentimen positif tercatat. Sementara itu, Keamanan Data dan Dampak lebih banyak didominasi sentimen netral, meskipun tetap disertai proporsi negatif yang signifikan. Sentimen positif secara umum sangat rendah di semua kategori, yang mencerminkan kecenderungan publik untuk mengkritik atau bersikap skeptis terhadap penanganan insiden PDNS. Distribusi pelabelan data berdasarkan aspek secara lengkap disajikan pada Tabel 5 dan Gambar 3.

Tabel 5. Distribusi Hasil Pelabelan Data Berdasarkan Kategori

	Keamanan Data	Lembaga	Infrastruktur	Politik	Dampak
Positif	47	7	46	0	2
Negatif	486	1052	1111	524	441
Netral	557	428	463	183	208

f. Pembagian Dataset dan Data Latih
Pembagian dataset dibagi menjadi data latih dan data uji. Pengambilan rasio ini didasarkan pada penelitian Mohammed Shenify tahun 2023 [40]. Pembagian data latih dan data uji ini dapat dilihat pada Tabel 6.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 960−971
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7473

Tabel 6. Pembagian Data Latih dan Data Uji

Data Latih	Data Uji
90%	10%
80%	20%
70%	30%

3.2 *Assess*

a. *Random Forest* dan *Support Vector Machine*

Dalam penerapan algoritma *Random Forest*, hasil evaluasi menunjukkan bahwa performa model meningkat seiring dengan proporsi data uji yang lebih besar. Akurasi yang diperoleh adalah sebesar 89% pada rasio data latih dan data uji 90:10, kemudian meningkat menjadi 94% pada rasio 80:20, dan mencapai 96% pada rasio 70:30. Nilai presisi, *recall*, dan *F1-score* pada setiap rasio juga konsisten berada pada angka yang sama dengan akurasi, yaitu masing-masing 89%, 94%, dan 96%, menunjukkan kinerja yang stabil dan andal dari algoritma *Random Forest* dalam klasifikasi sentimen berbasis aspek.

Sementara itu, pada penerapan algoritma *Support Vector Machine* (SVM), akurasi yang diperoleh sebesar 85% pada rasio 90:10, kemudian meningkat menjadi 86% pada rasio 80:20, dan mencapai 89% pada rasio 70:30. Nilai presisi, *recall*, dan *F1-score* juga menunjukkan peningkatan bertahap, yaitu dari 87%, 89%, dan 90% untuk presisi; serta 85%, 86%, dan 89% untuk *recall* dan *F1-score* secara berurutan. Hasil ini menunjukkan bahwa performa SVM meningkat seiring dengan bertambahnya data uji, meskipun tidak sekuat Random Forest. Evaluasi lengkap dapat dilihat pada Tabel 7.

Tabel 7. Evaluasi Performa *Random Forest* dan *Support Vector Machine*

	Random Forest				*Support Vector Machine*			
Data Latih: Data Uji	Akurasi	Presisi	*Recall*	*F1-Score*	Akurasi	Presisi	*Recall*	*F1-Score*
90:10	89%	89%	89%	89%	85%	87%	85%	85%
80:20	94%	94%	94%	94%	86%	89%	86%	86%
70:30	96%	96%	96%	96%	89%	90%	89%	89%

b. Metode SMOTE

Distribusi kelas pada Tabel 4 menunjukkan adanya ketidakseimbangan antar kelas pelabelan. Meskipun hasil evaluasi awal algoritma *Random Forest* dan *Support Vector Machine* (Tabel 7) menunjukkan performa model yang cukup tinggi, namun masih terdapat potensi bias terhadap kelas mayoritas. Pada algoritma *Random Forest*, akurasi berkisar antara 89% hingga 96%, sedangkan pada SVM, akurasi relatif lebih rendah dan bervariasi dari 85% hingga 89%, dengan ketidakseimbangan antara presisi dan *recall*. Untuk mengatasi masalah ketidakseimbangan data, diterapkan metode SMOTE pada algoritma *Random Forest* dan *Support Vector Machine* untuk meningkatkan kemampuan generalisasi model terhadap kelas minoritas. Hasil penerapan SMOTE pada distribusi sentimen disajikan dalam Tabel 8.

Tabel 8. Hasil Penerapan SMOTE

	Data Latih : Data Uji	Negatif	Netral	Positif
Sebelum SMOTE	90:10	361	184	10
	80:20	723	368	20
	70:30	1084	552	30
Setelah menggunakan SMOTE	90:10	361	361	361
	80:20	723	723	723
	70:30	1084	1084	1084

Tabel 9. Evaluasi Performa *Random Forest* dan *Support Vector Machine* Setealah menggunakan SMOTE

	Random Forest				*Support Vector Machine*			
Data Latih : Data Uji	Akurasi	Presisi	Recall	F1-Score	Akurasi	Presisi	Recall	F1-Score
90:10	93%	93%	93%	93%	91%	92%	91%	91%
80:20	95%	95%	95%	95%	95%	95%	95%	95%
70:30	96%	96%	96%	96%	95%	95%	95%	95%

Penerapan teknik SMOTE terbukti secara signifikan meningkatkan performa model baik pada algoritma Random Forest maupun Support Vector Machine (SVM), sebagaimana ditunjukkan pada Tabel 9. Pada algoritma Random Forest, akurasi mengalami peningkatan di seluruh rasio data latih dan uji, yaitu dari 89% menjadi 93% pada rasio 90:10, dari 94% menjadi 95% pada rasio 80:20, dan tetap tinggi di angka 96% pada rasio 70:30. Nilai presisi, recall, dan F1-score juga menunjukkan konsistensi yang sangat baik dengan angka yang sama seperti akurasi pada setiap rasio, menandakan kestabilan performa model setelah penerapan SMOTE.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 960−971
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7473

Pada algoritma SVM, dampak SMOTE juga terlihat jelas dengan peningkatan akurasi dari 85% menjadi 91% pada rasio 90:10, dari 86% menjadi 95% pada rasio 80:20, dan dari 89% menjadi 95% pada rasio 70:30. Seluruh metrik evaluasi seperti presisi, recall, dan F1-score juga menunjukkan peningkatan dan konsistensi di masing-masing rasio. Hasil ini menunjukkan bahwa SMOTE efektif dalam mengatasi ketidakseimbangan kelas, sehingga model dapat mengenali pola dari kelas minoritas dengan lebih baik dan menghasilkan klasifikasi sentimen berbasis aspek yang lebih akurat dan seimbang.

Penerapan teknik SMOTE secara signifikan meningkatkan performa model algoritma *Random Forest* dan *Support Vector Machine* (SVM), seperti yang ditunjukkan pada Tabel 9. Pada *Random Forest*, akurasi model menunjukkan peningkatan konsisten di berbagai rasio data latih dan uji, di mana akurasi naik dari 89% menjadi 91% (rasio 90:10), dari 93% menjadi 94% (rasio 80:20), dan tetap stabil di angka 96% (rasio 70:30). Sementara itu, SMOTE juga meningkatkan stabilitas dan keseimbangan performa algoritma SVM dengan akurasi model setelah SMOTE menjadi lebih konsisten, yaitu 88% pada rasio 90:10, 87% pada rasio 80:20, dan 88% pada rasio 70:30. Ini menunjukkan bahwa SMOTE efektif dalam menyeimbangkan representasi antar kelas, membantu kedua model mengenali pola dari kelas minoritas dengan lebih baik, dan pada akhirnya memperkuat akurasi klasifikasi secara keseluruhan.

c. *K-Fold Cross Validation*

Dalam menguji konsistensi kinerja model, dilakukan evaluasi menggunakan 10-*Fold Cross Validation*. Evaluasi dilakukan pada algoritma *Random Forest* dan SVM untuk membandingkan akurasi keduanya secara adil dalam kondisi data seimbang. Hasil lengkap disajikan pada Tabel 10.

Tabel 10. Perbandingan Hasil 10-*Fold Cross validation*

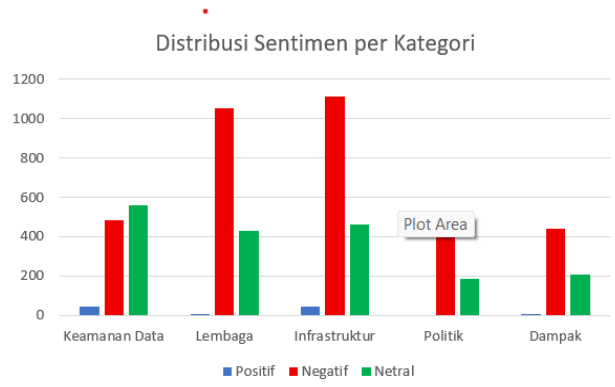
Fold	Akurasi Random Forest	Akurasi Support Vector Machine
1	99.10%	98.38%
2	99.10%	98.74%
3	98.02%	98.02%
4	98.02%	97.84%
5	99.10%	98.92%
6	99.10%	98.92%
7	99.27%	99.09%
8	99.63%	99.27%
9	98.73%	98.73%
10	99.27%	99.27%
Rata-rata	98.93%	98.82%

Tabel 10 menunjukkan bahwa *Random Forest* memiliki akurasi rata-rata sebesar 99.05%, sedikit lebih tinggi dibandingkan SVM yang mencapai rata-rata 98.86%. Performa tinggi dan konsisten ini menunjukkan bahwa kedua model dapat bekerja secara efektif. Kinerja yang stabil antar *fold* juga menunjukkan bahwa model tidak mengalami *overfitting* dan mampu melakukan klasifikasi dengan baik terhadap seluruh kategori sentimen.

3.3 Tahap Analisis Data

a. Analisi Klasifikasi

Klasifikasi sentimen terhadap insiden kebocoran Pusat Data Nasional Sementara (PDNS) dilakukan berdasarkan lima kategori aspek utama: Keamanan Data, Tokoh dan Lembaga, Infrastruktur, Politik dan Ekonomi, serta Dampak. Pelabelan dilakukan menggunakan model *pre-trained* IndoBERTweet, dengan proses pengelompokan aspek melalui POS Tagging, NER, TF-IDF, dan *semantic coherence*.



Gambar 3. Visualisasi Distribusi Sentimen per Kategori

Berdasarkan hasil pelabelan terhadap 2.700 tweet, sentimen negatif mendominasi hampir seluruh aspek. Pada aspek Infrastruktur, terdapat 1.111 tweet negatif, jauh melampaui 463 tweet netral dan hanya 46 tweet positif,

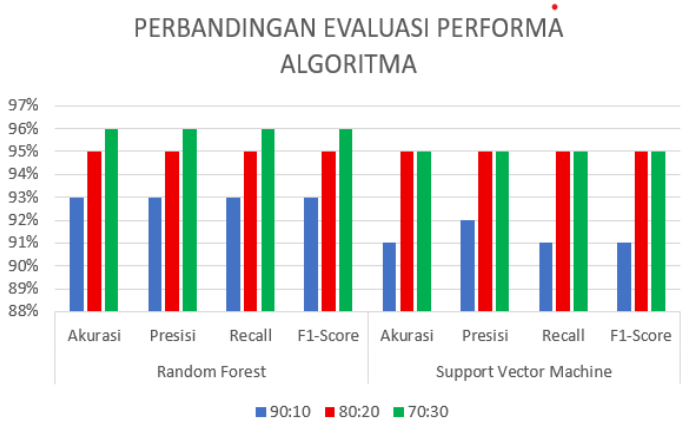
Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 960−971
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7473

menunjukkan kekecewaan publik terhadap kualitas infrastruktur digital pemerintah. Aspek Lembaga Pemerintah juga menunjukkan dominasi sentimen negatif yang kuat dengan 1.052 tweet negatif, dibanding 428 netral dan hanya 7 positif, mencerminkan kritik tajam terhadap kinerja serta akuntabilitas institusi terkait.

Aspek Keamanan Data menjadi satu-satunya yang didominasi oleh sentimen netral, yaitu sebanyak 557 tweet, dengan 486 negatif dan 47 positif. Hal ini mengindikasikan bahwa masyarakat cenderung menyampaikan ulang informasi tanpa respons emosional yang ekstrem, meskipun kekhawatiran tetap terlihat dari jumlah sentimen negatif yang cukup besar. Sementara itu, pada aspek Politik, tidak terdapat tweet positif sama sekali; sentimen negatif mencapai 524 tweet, dan 183 tweet bersifat netral, menunjukkan rendahnya kepercayaan publik terhadap pengaruh politik dalam isu ini.

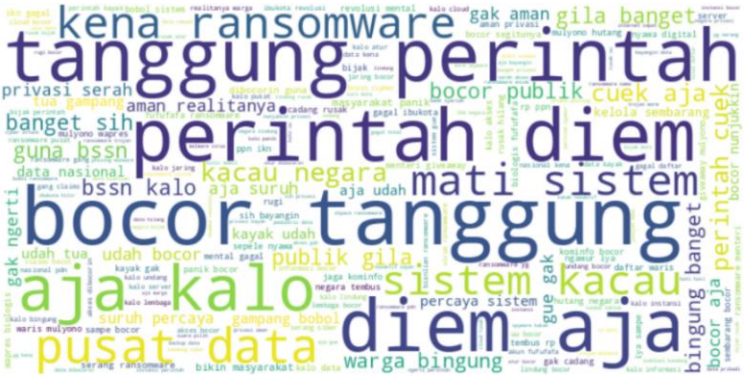
Pada aspek Dampak, terdapat 441 tweet negatif, 208 netral, dan hanya 2 tweet positif, menandakan bahwa masyarakat masih bersikap hati-hati dan menunggu perkembangan lebih lanjut terkait dampak nyata dari insiden kebocoran data tersebut.

Untuk mendukung klasifikasi ini secara teknis, dilakukan evaluasi model menggunakan dua algoritma pembelajaran mesin, yaitu *Random Forest* dan *Support Vector Machine* (SVM). Kedua algoritma diuji pada tiga rasio pembagian data (90:10, 80:20, dan 70:30) serta divalidasi menggunakan teknik SMOTE dan 10-*fold cross validation*.



Gambar 4. Perbandingan Evaluasi Performa *Random Forest* dan *Support Vector Machine*

Hasil evaluasi menunjukkan bahwa Random Forest secara konsisten unggul dalam metrik akurasi, presisi, recall, dan F1-score dibandingkan SVM. Pada rasio data latih dan uji 70:30, Random Forest mencapai akurasi 96%, sedangkan SVM memperoleh 95%. Keunggulan ini semakin ditegaskan melalui hasil 10-fold cross-validation, di mana Random Forest mencatat rata-rata akurasi 99,12%, sedikit lebih tinggi dibandingkan SVM yang mencapai 98,89%. Meskipun evaluasi performa algoritma bukan fokus utama penelitian ini, validasi tersebut memperkuat kualitas hasil klasifikasi sentimen yang telah dilakukan. Sebagai pelengkap visualisasi, ditampilkan pula word cloud untuk menggambarkan kata-kata dominan dalam opini publik terkait insiden PDNS.



Gambar 5. Visualisasi *Wordcloud* Negatif

Kata-kata seperti “gagal”, “retas”, “kominfo”, dan “virus” muncul secara signifikan, memperkuat ekspresi kekecewaan publik terhadap kelemahan sistem serta Lembaga Pemerintah pengelola data. *Wordcloud* ini berfungsi sebagai ilustrasi atas narasi dan statistik yang telah dikemukakan sebelumnya.

Dengan demikian, analisis klasifikasi ini berhasil mengungkap persepsi masyarakat per aspek dengan pendekatan ABSA, yang memperlihatkan fokus ketidakpuasan terhadap kelembagaan dan infrastruktur. Evaluasi model juga memperkuat keandalan hasil klasifikasi, menjadikan temuan ini relevan untuk menjadi dasar kebijakan publik di bidang keamanan siber nasional.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 960−971
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7473

b. Perbandingan dengan Penelitian Terdahulu
Untuk mengevaluasi efektivitas pendekatan yang diusulkan dalam penelitian ini, dilakukan analisis komparatif terhadap sejumlah studi terdahulu yang mengkaji isu serupa, khususnya terkait analisis sentimen terhadap fenomena kebocoran data. Tabel 11 menyajikan ringkasan perbandingan tersebut berdasarkan karakteristik data yang digunakan, metode klasifikasi yang diterapkan, serta hasil evaluasi performa model pada masing-masing penelitian.

Tabel 11. Perbandingan dengan Penelitian Terdahulu

No	Penulis (Tahun)	Metode	Akurasi/F1-score	Catatan
1	Sholehurrohman & Naïve Bayes (NB) Ilman (2022)		83.06%	Tidak menggunakan ABSA
2	Turmudi & NB, SVM Hadikristanto (2023)		NB: 97%, SVM: 80%	Tidak dijelaskan balancing data
3	Taufan & Wibowo (2024)	RF, SVM	SVM: 92.6%, AUC: 97%	Fokus kebijakan, bukan ABSA
4	Amal et al. (2022)	SVM, RF, IndoBERT, LR	SVM: F1 0.81, RF: 0.78	IndoBERT kurang optimal (F1 0.76)
5	Penelitian Ini	RF, SVM + ABSA	RF: 96.38%, SVM: 95.61%	Menggunakan IndoBERTweet + POS Tag + NER + TF-IDF + SMOTE

Berdasarkan Tabel 11, pendekatan yang diterapkan dalam penelitian ini menunjukkan kinerja yang kompetitif, bahkan dalam beberapa aspek melampaui capaian studi terdahulu. Capaian ini terutama didukung oleh integrasi teknik ABSA, pemanfaatan model IndoBERTweet yang kontekstual terhadap bahasa media sosial Indonesia, serta penerapan SMOTE untuk mengatasi ketidakseimbangan kelas. Kombinasi strategi ini berkontribusi pada peningkatan akurasi dan kedalaman pemahaman model terhadap sentimen publik dalam konteks kebocoran data PDNS.

4 KESIMPULAN

Kajian ini menganalisis opini publik terhadap insiden kebocoran Pusat Data Nasional Sementara (PDNS) dengan menggunakan pendekatan *Aspect-Based Sentiment Analysis* (ABSA), yang memetakan sentimen secara tematik berdasarkan lima kategori utama: Keamanan Data, Lembaga, Infrastruktur, Politik, dan Dampak. Pelabelan sentimen dilakukan melalui model IndoBERTweet, dengan proses evaluasi menggunakan algoritma *Random Forest* dan *Support Vector Machine* (SVM) untuk menjamin keandalan hasil klasifikasi. Temuan menunjukkan dominasi sentimen negatif di hampir seluruh kategori, terutama pada kategori Lembaga Pemerintah dan Infrastruktur, yang mengindikasikan kekecewaan publik terhadap akuntabilitas institusi dan ketahanan sistem digital nasional. Di sisi lain, sentimen netral pada kategori Keamanan Data mencerminkan sikap informatif dan kehati-hatian masyarakat. Ketegangan opini juga muncul pada kategori Politik dan Dampak, menandakan adanya kekhawatiran terhadap konsekuensi kebijakan dan risiko jangka panjang dari insiden ini. Secara umum, pendekatan ABSA terbukti efektif dalam menangkap dinamika opini publik secara terstruktur dan kontekstual. Sedangkan pada hasil evaluasi performa dari dua algoritma yang digunakan, *Random Forest* memberikan performa paling optimal, dengan akurasi dan stabilitas klasifikasi yang konsisten tinggi di berbagai skenario pengujian. Hal ini memperkuat keandalan pendekatan ABSA yang diterapkan, serta mendukung validitas hasil pelabelan sentimen terhadap masing-masing aspek yang dianalisis. Meskipun telah memberikan gambaran yang komprehensif, kajian ini memiliki keterbatasan pada ruang lingkup data yang terbatas pada platform X dan kurun waktu tertentu. Selain itu, model klasifikasi berbasis *supervised learning* yang digunakan belum sepenuhnya mampu menangkap ekspresi emosional kompleks seperti sarkasme atau ironi. Untuk itu, pengembangan model lintas platform serta pendekatan berbasis emosi dan konteks multimodal menjadi arah yang layak dipertimbangkan untuk penelitian lanjutan.

REFERENCES

[1] H. T. Halawani, A. M. Mashraqi, S. K. Badr, and S. Alkhalaf, "Automated sentiment analysis in social media using Harris Hawks optimisation and deep learning techniques," *Alexandria Engineering Journal*, vol. 80, pp. 433–443, Oct. 2023, doi: 10.1016/j.aej.2023.08.062.

[2] H. Alhindi, I. Traore, and I. Woungang, "Preventing Data Leak through Semantic Analysis," *Internet of Things*, vol. 14, p. 100073, Jun. 2021, doi: 10.1016/j.iot.2019.100073.

[3] D. Arisandi, T. Sutrisno, and I. Kurniawan, "KLASIFIKASI OPINI MASYARAKAT DI TWITTER TENTANG KEBOCORAN DATA YANG TERJADI DI INDONESIA MENGGUNAKAN ALGORITMA SVM," *Jurnal Informatika Kaputama (JIK)*, vol. 7, no. 1, pp. 84–90, Jan. 2023, doi: 10.59697/jik.v7i1.10.

[4] Z. N. Aziza and D. Y. Kristiyanto, "Prediction of The Level of Public Trust in Government Policies in the 1 st Quarter of The Covid 19 Pandemic using Sentiment Analysis," *E3S Web of Conferences*, vol. 317, p. 05013, Nov. 2021, doi: 10.1051/e3sconf/202131705013.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 960−971
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7473

[5] V. Tandon and R. Mehra, "An Integrated Approach For Analysing Sentiments On Social Media," *Informatica*, vol. 47, no. 2, pp. 213–220, Jun. 2023, doi: 10.31449/inf.v47i2.4390.
[6] Z. Janková, "CRITICAL REVIEW OF TEXT MINING AND SENTIMENT ANALYSIS FOR STOCK MARKET PREDICTION," *Journal of Business Economics and Management*, vol. 24, no. 1, pp. 177–198, Apr. 2023, doi: 10.3846/jbem.2023.18805.
[7] P. Guo, "Construction of Semantic Coherence Diagnosis Model of English Text based on Sentence Semantic Map," *Scalable Computing: Practice and Experience*, vol. 25, no. 1, pp. 327–339, Jan. 2024, doi: 10.12694/scpe.v25i1.2298.
[8] H.-H. Nguyen, "Enhancing Sentiment Analysis on Social Media Data with Advanced Deep Learning Techniques," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 5, 2024, doi: 10.14569/IJACSA.2024.0150598.
[9] T. Ahmed Khan, R. Sadiq, Z. Shahid, M. M. Alam, and M. Mohd Su'ud, "Sentiment Analysis using Support Vector Machine and Random Forest," *Journal of Informatics and Web Engineering*, vol. 3, no. 1, pp. 67–75, Feb. 2024, doi: 10.33093/jiwe.2024.3.1.5.
[10] N. I. Wibowo, T. A. Maulana, H. Muhammad, and N. A. Rakhmawati, "Perbandingan Algoritma Klasifikasi Sentimen Twitter Terhadap Insiden Kebocoran Data Tokopedia," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 6, no. 2, pp. 120–129, May 2021, doi: 10.14421/jiska.2021.6.2.120-129.
[11] R. Sholehurrohman and I. Sabda Ilman, "ANALISIS SENTIMEN TWEET KASUS KEBOCORAN DATA PENGGUNAAN FACEBOOK OLEH CAMBRIGDE ANALYTICA," *Jurnal Pepadun*, vol. 3, no. 1, pp. 140–147, Apr. 2022, doi: 10.23960/pepadun.v3i1.108.
[12] A. Zy and Wahyu Hadikristanto, "Implementasi Algoritma Metode Naive Bayes dan Support Vector Machine Tentang Pembobolan dan Kebocoran Data di Twitter," *Bulletin of Information Technology (BIT)*, vol. 4, no. 1, pp. 49–56, Mar. 2023, doi: 10.47065/bit.v4i1.493.
[13] C. Umam, L. B. Handoko, and F. O. Isinkaye, "Performance Analysis of Support Vector Classification and Random Forest in Phishing Email Classification," *Scientific Journal of Informatics*, vol. 11, no. 2, pp. 367–374, May 2024, doi: 10.15294/sji.v11i2.3301.
[14] A. Z. Taufan and W. Wibowo, "ANALISIS SENTIMEN TERKAIT PERSEPSI KEAMANAN DATA INFORMASI DAN PRIVASI DI INDONESIA MENGGUNAKAN PENDEKATAN MACHINE LEARNING," *Jurnal Informatika Teknologi dan Sains (Jinteks)*, vol. 6, no. 3, pp. 728–736, Aug. 2024, doi: 10.51401/jinteks.v6i3.4764.
[15] M. I. Amal, E. S. Rahmasita, E. Suryaputra, and N. A. Rakhmawati, "Analisis Klasifikasi Sentimen Terhadap Isu Kebocoran Data Kartu Identitas Ponsel di Twitter," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, no. 3, Dec. 2022, doi: 10.28932/jutisi.v8i3.5483.
[16] A. M. Taufiqi and A. Nugroho, "Sentimen Pengguna Twitter Mengenai Isu Kebocoran Data Dengan Algoritma Naïve Bayes," *Jurnal Nasional Ilmu Komputer*, vol. 4, no. 1, pp. 1–11, Mar. 2023, doi: 10.47747/jurnalnik.v4i1.1091.
[17] D. Jacob and R. Henriques, "Educational Data Mining to Predict Bachelors Students' Success," *Emerging Science Journal*, vol. 7, no. Special Issue 2, pp. 159–171, Jul. 2023, doi: 10.28991/ESJ-2023-SIED2-013.
[18] L. Andrade-Arenas, I. Rubio-Paucar, and C. Yactayo-Arias, "Predictive models in Alzheimer's disease: an evaluation based on data mining techniques," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 3, p. 2988, Jun. 2024, doi: 10.11591/ijece.v14i3.pp2988-3002.
[19] J. Boegershausen, H. Datta, A. Borah, and A. T. Stephen, "Fields of Gold: Scraping Web Data for Marketing Insights," *J Mark*, vol. 86, no. 5, pp. 1–20, Sep. 2022, doi: 10.1177/00222429221100750.
[20] V. Boppana and P. Sandhya, "Web crawling based context aware recommender system using optimized deep recurrent neural network," *J Big Data*, vol. 8, no. 1, p. 144, Dec. 2021, doi: 10.1186/s40537-021-00534-7.
[21] N. Babanejad, H. Davoudi, A. Agrawal, A. An, and M. Papagelis, "The Role of Preprocessing for Word Representation Learning in Affective Tasks," *IEEE Trans Affect Comput*, vol. 15, no. 1, pp. 254–272, Jan. 2024, doi: 10.1109/TAFFC.2023.3270115.
[22] C. B. Lee, H. N. Io, and H. Tang, "Sentiments and perceptions after a privacy breach incident," *Cogent Business and Management*, vol. 9, no. 1, 2022, doi: 10.1080/23311975.2022.2050018.
[23] P. A. Rodríguez-Correa *et al.*, "Information security education: a thematic trend analysis," *F1000Res*, vol. 14, p. 5, Jan. 2025, doi: 10.12688/f1000research.159828.1.
[24] R. Shandler, N. Kostyuk, and H. Oppenheimer, "Public Opinion and Cyberterrorism," *Public Opin Q*, vol. 87, no. 1, pp. 92–119, 2023, doi: 10.1093/poq/nfad006.
[25] N. U. Prince *et al.*, "AI-Powered Data-Driven Cybersecurity Techniques: Boosting Threat Identification and Reaction," 2024. [Online]. Available: www.nano-ntp.com
[26] M. Sivakumar, S. Parthasarathy, and T. Padmapriya, "Trade-off between training and testing ratio in machine learning for medical image processing," *PeerJ Comput Sci*, vol. 10, p. e2245, Sep. 2024, doi: 10.7717/peerj-cs.2245.
[27] V. M. Rajan and A. Ramanujan, "Architecture of a Semantic WordCloud Visualization," in *Second International Conference on Networks and Advances in Computational Technologies*, L. and J. J. and J. J. Palesi Maurizio and Trajkovic, Ed., Cham: Springer International Publishing, 2021, pp. 95–106. doi: 10.1007/978-3-030-49500-8_9.
[28] M. S. Sayeed, V. Mohan, and K. S. Muthu, "BERT: A Review of Applications in Sentiment Analysis," *HighTech and Innovation Journal*, vol. 4, no. 2, pp. 453–462, Jun. 2023, doi: 10.28991/HIJ-2023-04-02-015.
[29] A. F. Hidayatullah, R. A. Apong, D. T. C. Lai, and A. Qazi, "Corpus creation and language identification for code-mixed Indonesian-Javanese-English Tweets," *PeerJ Comput Sci*, vol. 9, p. e1312, Jun. 2023, doi: 10.7717/peerj-cs.1312.
[30] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 10660–10668. doi: 10.18653/v1/2021.emnlp-main.833.
[31] M. Kurniawan, K. Kusrini, and M. R. Arief, "Part of Speech Tagging Pada Teks Bahasa Indonesia dengan BiLSTM + CNN + CRF dan ELMo," *Jurnal Eksplora Informatika*, vol. 11, no. 1, pp. 29–37, Jan. 2022, doi: 10.30864/eksplora.v11i1.506.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 960−971
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7473

[32] E. Yulianti, N. Bhary, J. Abdurrohman, F. W. Dwitilas, E. Q. Nuranti, and H. S. Husin, "Named entity recognition on Indonesian legal documents: a dataset and study using transformer-based models," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 5, p. 5489, Oct. 2024, doi: 10.11591/ijece.v14i5.pp5489-5501.
[33] L. Wang, Y. Yang, L. Xu, and T. Ji, "Application of random forest algorithm in the detection of foreign objects in wine," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, Jan. 2024, doi: 10.2478/amns.2023.2.00055.
[34] D. Papakyriakou and I. S. Barbounakis, "Data Mining Methods: A Review," *Int J Comput Appl*, vol. 183, no. 48, pp. 5–19, Jan. 2022, doi: 10.5120/ijca2022921884.
[35] V. D. Cong and T. T. Hiep, "Support vector machine-based object classification for robot arm system," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 5, p. 5047, Oct. 2023, doi: 10.11591/ijece.v13i5.pp5047-5053.
[36] V. Ganganwar and R. Rajalakshmi, "Employing synthetic data for addressing the class imbalance in aspect-based sentiment classification," *Journal of Information and Telecommunication*, vol. 8, no. 2, pp. 167–188, Apr. 2024, doi: 10.1080/24751839.2023.2270824.
[37] A. Newaz, Md. S. Mohosheu, Md. A. Al Noman, and T. Jabid, "iBRF: Improved Balanced Random Forest Classifier," in *2024 35th Conference of Open Innovations Association (FRUCT)*, Tampere: IEEE, Apr. 2024, pp. 501–508. doi: 10.23919/FRUCT61870.2024.10516372.
[38] M. F. Schrauf, G. de los Campos, and S. Munilla, "Comparing Genomic Prediction Models by Means of Cross Validation," *Front Plant Sci*, vol. 12, Nov. 2021, doi: 10.3389/fpls.2021.734512.
[39] J. Wieczorek, C. Guerin, and T. McMahon, "*K* -fold cross-validation for complex sample surveys," *Stat*, vol. 11, no. 1, Dec. 2022, doi: 10.1002/sta4.454.
[40] M. Shenify, "Sentiment analysis of Saudi e-commerce using naïve bayes algorithm and support vector machine," *International Journal of Data and Network Science*, vol. 8, no. 3, pp. 1607–1612, Jun. 2024, doi: 10.5267/j.ijdns.2024.3.006.